Impacts of Performance Pay Under the Teacher Incentive Fund: Study Design Report

October 2011

Steven Glazerman
Hanley Chiang
Alison Wellington
Jill Constantine
Dan Player

MATHEMATICA
Policy Research

Contract Number:
ED-04-CO-0112-0012

Mathematica Reference Number:
06715.500

Submitted to:
Institute of Education Sciences
555 New Jersey Avenue, NW
Room 502A
Washington, DC 20208-5500
Telephone: (202) 208-7169
Project Officer: Elizabeth Warner

Submitted by:
Mathematica Policy Research
600 Maryland Avenue, SW
Suite 550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Jill Constantine

Impacts of Performance Pay Under the Teacher Incentive Fund: Study Design Report

October 2011

Steven Glazerman
Hanley Chiang
Alison Wellington
Jill Constantine
Dan Player

MATHEMATICA
Policy Research

CONTENTS

TABLES

FIGURES

I. INTRODUCTION

A large and growing body of research reveals that teacher quality is a critical input to student learning (Hanushek 2010; Gordon et al. 2006; Rivkin et al. 2005; Rockoff 2004). A highly effective teacher can have a significant influence on student achievement gains, with impacts accumulating over time for students consistently taught by effective teachers (Hanushek et al. 2005; Jordan et al. 1997; Sanders and Rivers 1996).

Although it is important for all schools to have high quality teachers, evidence suggests that the problem of attracting and retaining these educators is even more pronounced for high-need schools and hard-to-staff subjects (Glazerman and Max 2011; Boyd et al. 2008; Jacob 2007; Monk 2007; Tennessee Department of Education 2007; Iatarola and Stiefel 2003; Kirby et al. 1999; Jordan et al. 1997; Sanders and Rivers 1996). Furthermore, little is known about how to develop a strong teacher workforce (Rivkin et al. 2005; Rockoff 2004).

Under the uniform salary schedule, which has been the dominant compensation structure for teachers in the United States since the 1920s, teachers' salaries are driven by two primary factors: (1) tenure in the classroom and (2) completed postgraduate work. However, the research does not suggest a strong relationship between these factors (experience and coursework/degrees) and student achievement growth. Research from school districts in Chicago, Florida, and Texas suggests that length of time in the teaching profession matters only during the first two to three years in the classroom (Harris and Sass 2008; Aaronson et al. 2007; Hanushek et al. 2005). Thereafter, less-experienced teachers might be as effective as longer-term veterans. The same research also indicates that advanced degrees generally do not help distinguish between more- and less-effective teachers.

A. Pay- for- Performance Initiatives and Evidence

Given that pay under the traditional salary schedule is not based on performance, recent efforts to reform teacher compensation have focused on how we evaluate and compensate teachers. Some compensation reforms are designed to evaluate teacher effectiveness by estimating teachers' contributions to students' achievement gains and then rewarding teachers, in part, on this dimension of performance. As discussed in the conceptual framework in Section C, one hypothesis maintains that rewarding teachers for student achievement gains will improve students' achievement by attracting more effective teachers to the field or improving the effectiveness of existing teachers.

Efforts during the 1980s and 1990s to reform the uniform salary schedule—including practices such as performance-based pay, responsibility pay (career ladders), and bonuses for teaching in high-need subjects or geographic areas—resulted in program implementation of short duration and with a variety of implementation challenges (Glazerman 2004). Many researchers have examined these implementation difficulties (Silman and Glazerman 2009; Glazerman 2004; Podgursky 2002; Hatry et al. 1994; Murnane and Cohen 1986) and attribute the lack of implementation success to several factors, including the challenge of developing objective and reliable measures of teacher performance, stakeholder reluctance to support differential pay for teachers, and the cost of sustaining the programs financially.

More recently, the data and technology for measuring teacher performance have advanced. The No Child Left Behind Act increased the frequency and coverage of standardized student assessment, development of sophisticated student data-tracking systems, and the advancement of methods for computing a teacher's contribution to his or her students' achievement growth or "value-added"

indicators of teachers' performance. At the same time, political and financial support for teacher compensation reform has grown at all levels—federal, state, and local. For instance, the U.S. Department of Education (ED) awarded $95 million in 2007 to 34 grantees through the Teacher Incentive Fund (TIF). State and local initiatives include Minnesota's Q-Comp program ($86 million), Nevada's teacher bonus programs ($65 million), and New York City's Partnership for Teaching Excellence ($15 million granted by the Carrol and Milton Petrie Foundation), and a multisite investment of $370 million by the Bill & Melinda Gates Foundation to reform teacher compensation in Pittsburgh, Pennsylvania; Memphis, Tennessee; Hillsborough County, Florida; and a consortium of charter schools in California.

A recent review of research suggested mixed evidence on the ability of teacher pay reform to improve student achievement and teacher retention. Podgursky and Springer (2007) reviewed 10 studies of performance-based teacher pay and reported that most of the studies found a positive association between the incentive program and student achievement. However, for several reasons, these studies provided only limited insight into the impact that incentive programs can have on student achievement in the United States. First, neither of the two most rigorous studies, which used random assignment designs and had positive findings, occurred in the United States (one was in rural India and the other in Kenya). Second, in the five studies conducted in the United States, some examined programs that were implemented in all schools in the district or in only one school, making it difficult to find appropriate comparisons. Thus, in the studies in the United States, we cannot be confident that the differences in student achievement were due to the teacher pay programs rather than some other factor about the schools or districts in which the policy was implemented.[1]

Performance-based compensation programs have expanded rapidly in the past decade and have been piloted in several large states and school districts. Several studies have been completed since the publication of the Podgursky and Springer review (2007), with large samples of students and teachers, including more-rigorous studies based on random assignment study designs. In the following pages, we describe the findings on some of the more prevalent models of performance-based compensation, as well as studies that examined effects of the programs over several years.

Several studies have examined the Teacher Advancement Program (TAP), a comprehensive teacher pay reform model that, among its other features, provides performance-based pay. The TAP model includes evaluation based, in part, on student achievement and classroom observation scores and offers substantial salary supplements (a suggested range of $5,000 to $20,000 annually) for a few teachers who are selected to be mentor or master teachers. Mentor and master teachers perform additional duties and hold leadership roles in TAP schools. Not only are all teachers eligible for performance awards, they receive ongoing professional development and conduct weekly data-focused meetings called cluster groups to help them stay focused on academic goals.

The earliest studies of TAP, published or sponsored by the program's developers, reported positive associations between student achievement gains and TAP participation (Solmon et al. 2007; Schacter et al. 2004; Schacter et al. 2002). The authors compared student gains between TAP and selected non-TAP schools. However, these studies also had study designs in which schools and

[1] In addition, Podgursky and Springer (2007) raised several other methodological concerns about two of the studies.

districts that chose to implement TAP were compared with schools and districts that did not, making it difficult to disentangle the effects of TAP from initial characteristics of schools and districts. For example, in one study, authors reported that a higher percentage of TAP than non-TAP schools were in high achievement categories at the end of the year, but without controlling for initial achievement levels, we cannot reliably attribute those differences to TAP.

Independent studies that also compare schools that chose to implement TAP and those that did not showed mixed results. Springer et al. (2008) compared student test score gains in mathematics in TAP and non–TAP schools in two anonymous states. They used a panel data set that included approximately 1,200 schools over a four-year period, of which 28 schools implemented TAP at some point. The authors reported a positive association between student achievement and TAP for elementary students but negative findings for students in grades 6 through 10 after using statistical controls for initial student achievement in schools.

In another study of TAP, researchers examined the program's impact in Chicago Public Schools (CPS) using a hybrid study design, with both a randomized experiment and a matched comparison analysis. Chicago TAP provided substantial pay supplements for teachers who were selected to be mentors or master teachers ($7,000 and $15,000, respectively), but the payouts based purely on performance averaged $1,100 in the first year and $2,600 in the second year, or approximately 2 to 5 percent of base pay for a teacher earning $50,000. Furthermore, the maximum payouts were $2,045 and $6,320. This suggests that the highest-performing teachers did not earn substantially more than the average teacher, especially in the first year of the program. Sixteen CPS elementary schools volunteered to implement TAP, 8 of which were randomly chosen by the researchers to implement TAP in 2007; the other 8 were assigned to a control group that had to delay implementation of TAP until 2008. The remainder of the district's more than 300 schools were considered for inclusion in a complementary matched comparison analysis. The first-year impact results showed no impact of TAP on test scores in Chicago; however, the authors found evidence of a positive impact on teacher retention (Glazerman et al. 2009). The second-year follow-up, which extended the experimental analysis but relied more heavily on the matched comparison design, found no relationship between TAP and student achievement or teacher retention (Glazerman and Seifullah 2010). A final report, due to be released in 2012, is expected to provide longer-term results (four years) and include a much larger sample because Chicago Public Schools ultimately implemented its program in 40 schools.

Several studies have focused on pay-for-performance incentive programs in Texas, which began considering teacher pay reform during the 1980s (Springer et al. 2009a, 2009b). Texas has implemented several large-scale pay-for-performance incentive programs with funding from the state, federal government, or both. These include the Governor's Educator Excellence Grant (GEEG), Texas Educator Excellence Grant (TEEG), and District Awards for Teacher Excellence (DATE.).[2] Both GEEG and TEEG provided grants to implement pay-for-performance bonuses for teachers at schools with high or improved student achievement and serving a high percentage of economically disadvantaged students. Although both programs suggested minimum bonus awards of $3,000 and maximum awards of $10,000, these goals were typically not achieved. For instance, more than 80 percent of the schools participating in these programs proposed minimum bonuses of

[2] The first year of implementation of DATE was the 2008–2009 school year and to date there are no published findings on the impact of the program on student achievement or teacher turnover.

less than $3,000 and 45 to 80 percent proposed maximum awards of less than $3,000. Furthermore, the average difference between the minimum and maximum awards proposed by schools was less than $2,000, suggesting that awards for the highest-performing teachers were not substantially higher than for average teachers. Studies of TEEG and GEEG had a mixture of study designs that attempted to control for student achievement before the implementation of the program. The GEEG study examined changes in student achievement over time by comparing program and nonprogram schools (the authors noted that differences in the schools themselves, not the GEEG program, might have led to changes in student achievement). The TEEG study used a regression discontinuity design that attempted to control for differences in program and nonprogram schools. The year 3 evaluation reports of GEEG and TEEG found no systematic association between the two programs and student achievement or teacher turnover. The authors found mixed evidence on the association between student achievement gains and planned design features. However, both reports found that the receipt and size of the incentive bonus were strongly correlated with teacher turnover within the GEEG and TEEG schools, with the probability of turnover decreasing as the size of the bonus increased.

In the past two years, a number of multiyear studies have been conducted on pay-for-performance programs in Guilford, North Carolina; Nashville, Tennessee; and New York City. A summary of the programs and the research findings follows:

- Mission Possible, a TIF-funded program implemented in Guilford County Schools in North Carolina, is a comprehensive recruitment and retention program that included performance bonuses ranging between $2,500 and $5,000, as well as recruitment bonuses, professional development, performance accountability, and structural support. A study on the effect of Mission Possible on students' achievement (Bayonas 2010), compared outcomes of 28 Mission Possible schools with a matched comparison group of the other 70 district schools. The study examined student achievement for elementary through high school students. In general, differences between Mission Possible and comparison schools were not statistically significant.

- Three studies examined the impact of New York City's School-Wide Performance Bonus Program (SPBP), a program that was implemented in approximately 200 K–12 public schools in 2007–2008 and 2008–2009 (Marsh et al. 2011; Fryer 2011; Goodman and Turner 2010). Schools that met student achievement performance targets could earn up to $3,000 per full-time union member at the school. In each school, a school-level compensation committee decided how bonuses would be distributed and, in most cases, all teachers received the same amount of bonus when their school achieved the student performance target. The studies included a set of 402 low-performing, high-poverty schools, and 234 were randomly offered to participate in the program. The studies found no overall impact of SPBP on student achievement, high school graduation rates, or teacher retention or absences, and in a few cases, found small negative effects on math or reading achievement in certain years.

- Another recent study by Springer et al. (2010) examined the impact of the Project on Incentives in Teaching (POINT) program in Nashville, Tennessee. POINT offered substantial performance bonuses (ranging from $5,000 to $15,000) to middle school math teachers whose students achieved large gains on standardized tests, and the highest-performing teachers were eligible to receive bonuses substantially larger than the average teacher. Math teachers were given the opportunity to participate in a three-year experiment in which participants were randomly assigned to be eligible for the pay-for-

> performance bonuses or to the control condition, which was the usual salary schedule. The experiment included 296 teachers at the beginning of the first year, but there was substantial attrition from the sample and only 148 teachers participated by the third year. Substantial sample attrition makes it more difficult to attribute effects by year 3 to the program and not characteristics of the teachers and classrooms that remained in the study sample. The authors found no overall impact of performance bonuses on student math achievement, although there were positive impacts in years 2 and 3 on grade 5 students' achievement. The study did not examine the impact of performance pay awards on other important dimensions of current reform policies, such as teacher recruitment or whether incentives in combination with other reforms (such as pay for teachers to take on additional roles or responsibilities or targeted professional development) are associated with student achievement gains. The study also did not capture the potential importance of school-based incentives.

Although useful, the available evidence to date leaves policymakers with more questions than answers about the likely effects of a future performance pay program. First, would the findings hold up to more rigorous study methods? Relatively few studies rely on experimental designs that enable us to attribute differences in student achievement and teacher retention solely to the compensation program, and not other characteristics of the teachers, schools, or districts. Second, has the most comprehensive performance-based pay approach been put to the test? Even when the studies are well designed, the programs might be narrow or incomplete in some critical way. For example, evaluations that did not account for the effects of new teachers who entered after performance pay or did not measure program-induced attrition might fail to capture recruitment and retention effects. It is important to document whether programs studied included the program's initial and ongoing communication to teachers and principals about the performance metrics on which they would be evaluated or, alternatively, teacher's demonstrated understanding of and expectation of performance pay rules and payouts. Another key component might be targeted, data-driven professional development that could help teachers align their classroom practices to improve along the performance evaluation metric. To the extent that well-designed and well-implemented models are feasible in practice, evaluation efforts that focus on these experiences are most useful for informing policy. Finally, policymakers want to know if findings from selected sites can be generalized more broadly. Research should be conducted in a variety of settings within the United States, include sufficient numbers of schools and teachers, and follow up over a long enough period to support statistically reliable conclusions about the full effects of a performance pay system.

The body of research on the design, implementation, and effects of performance-based compensation systems has influenced the design and evaluation of the 2010 TIF grants. In the sections that follow, we describe the key components of 2010 TIF grants and the conceptual framework for the evaluation.

B. 2010 Teacher Incentive Fund Grants

The 2009 American Recovery and Reinvestment Act (ARRA) provides approximately $100 billion in funding to ED in support of programs to improve public education, particularly in high-need schools. TIF is an important component of this effort. With ARRA funds and ED's fiscal year 2010 appropriation, the department distributed more than $440 million in new federal TIF grants in support of performance-based teacher and principal compensation systems to reward and help attract and retain top education talent in high-need schools.

According to the grant requirements, 2010 TIF grantees must use their funds to provide the following:

- Performance pay that must include differentiated levels of compensation for teachers and principals based on effectiveness. Although grantees determined the specifics of their bonus programs, the grant notice states that the bonuses should be substantial and awarded based on challenging criteria. The notice also specifies that educator effectiveness must be based on student achievement growth and classroom or principal observations.
- Additional pay for educators who assume leadership roles and take on additional responsibilities.
- Targeted professional development.

In addition, grantees may offer a variety of other performance-based compensation incentives, such as additional pay to teach a hard-to-staff subject or in a high-need school.[3]

ARRA requires ED, to the extent possible, to conduct a rigorous national evaluation to assess the impact of performance-based compensation systems supported by ARRA funds, on student achievement and educator recruitment and retention in high-need schools and subjects. This evaluation, which meets this requirement, will provide important insights into performance pay programs using a randomized experiment as the study design, including multiple sites in the United States, targeting programs that are comprehensive, providing technical assistance to support strong implementation, and capturing all the hypothesized effects of such programs, as discussed next.

C. Conceptual Framework and Research Questions

Figure I.1 illustrates how incentive programs are hypothesized to affect student achievement. To have an impact, it is critical that educators are aware of the incentives, their eligibility for an incentive, and the criteria for receiving an incentive. Assuming that the program has been well communicated, incentives could, in theory, improve teacher effectiveness through two mechanisms: (1) a composition effect and (2) a productivity effect.

The composition effect could improve effectiveness of the educator workforce as a whole by attracting new, higher-performing teachers to schools that offer incentive programs and retaining higher-performing teachers. At the same time, this effect could encourage lower-performing teachers to leave their schools because they might feel stigmatized by not receiving an award and therefore would prefer to work in schools without incentives.

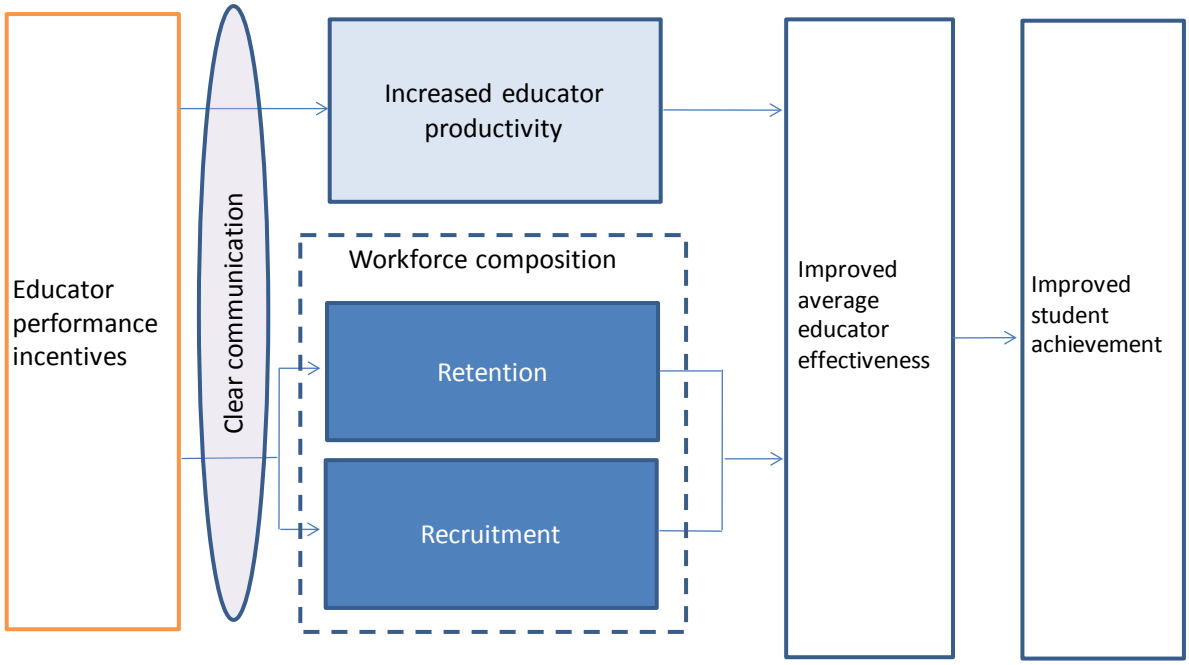
The productivity effect could improve effectiveness of individual educators by motivating teachers to improve their performance in order to receive incentive payments. This improvement

[3] For the purposes of this report, we will use the terms *performance pay* and *pay for performance* interchangeably to describe the component of the performance-based compensation system that rewards educators based *only* on their effectiveness. Although opportunities to take on additional roles and responsibilities might be limited to highly effective teachers, the additional pay for these roles is separate from what we refer to as performance pay and is independent of a teacher's ability to receive a performance pay award.

could result by increasing individual effort, increasing collaboration with colleagues, accessing professional development to align their teaching with measured outcomes, or some other method.

Teacher incentives could, however, have neutral or negative effects (Jacob and Levitt 2003; Jacob 2005). They could create the illusion of productivity gains by encouraging unintended educator behavior, such as teachers emphasizing test-taking strategies and not content knowledge. Incentives can also result in teachers allocating less time to nontested subjects, which also contribute to student learning but are more difficult to measure.

Figure I.1. Logic Model



There are several pathways in the logic model that could be blocked, preventing incentive programs from raising student achievement. If the program has not been communicated effectively, then educators have no motivation to change their behavior. It might take years of working within the system for the participants to fully understand both the evaluation criteria and the payout rules.

Even if the program is well communicated, other factors could impede the potential composition effects. For instance, if the district has a hiring freeze or a policy that discourages teacher transfers, we might not find the hypothesized composition effects. Even without such policy barriers, it could take years for the full composition effect to influence the workforce in a detectable way.

There could also be factors that prevent us from finding productivity effects. If the district or state has programs in place that already provide strong productivity incentives for educators, such as the threat of school closure or tenure requirements, then additional incentive programs might not produce additional productivity gains. In fact, it is possible that incentive programs might even reduce productivity if the program creates a negative school environment, perhaps by fostering competition and not collaboration among teachers.

This study will not only estimate the impact of performance pay on the final outcome of interest, student achievement, but will also examine intermediate outcomes, such as educators'

attitudes and awareness of the program and educators' mobility and recruitment, to provide a broader understanding of the mechanisms of how or why performance pay affects student achievement.

More specifically, the study will address the following research questions:

1. **What is the impact of performance-based pay on student achievement?** The evaluation will compare students' math and reading achievement in treatment and control schools, such that teachers in treatment schools can earn performance-based incentive bonuses. The estimated impacts on student achievement will reflect the combined impacts attributable to changes in the educator workforce and changes in educators' productivity.
2. **What is the impact of performance-based pay on educator mobility and recruitment?** As noted, a performance-based incentive program might affect student achievement through compositional or productivity effects. To provide insight into the mechanisms of how this might affect student achievement, the evaluation will also focus on the difference between educators' recruitment and retention in treatment versus control schools. To the extent possible, the evaluation will also examine how recruitment and retention differ by educator effectiveness.
3. **What are the experiences and challenges of grantees and districts when implementing performance-based compensation systems?** Understanding the implementation experiences and challenges of TIF grantees is crucial for interpreting impact findings and will provide essential information for improving the implementation of future incentive programs.
4. **Which types of performance pay models are associated with greater growth in student achievement and with educator mobility and retention?** To the extent possible, the study will examine the relative effectiveness of group-based (or mixed) or individual-based incentives programs.
5. **Do other important program features correlate with student and educator outcomes?** The study will examine how program features, such as the size and distribution of awards and the relative weights of student achievement growth and other measures of performance, are associated with student and educator outcomes.

D. Plan of Design Report

The remainder of this report describes our plans for the study design in greater detail. In Chapter II, we provide context for the 2010 TIF grants and describe the treatment and counterfactual conditions—that is, the conditions that would exist in the absence of the performance pay systems. In Chapter III, we discuss data collection, the sample for the evaluation, and how we will conduct random assignment. Finally, we explain our analysis plans in Chapter IV.

II. POLICY ENVIRONMENT: DEFINING THE TREATMENT AND ITS COUNTERFACTUAL

The purpose of the evaluation is to measure the impact of providing performance-based compensation for teachers and principals. However, policymakers often introduce performance pay at the same time that they change other aspects of teacher compensation and offer new types of professional development. This evaluation takes place in the context of TIF grants, which include several policy changes beyond performance-based compensation. We will study schools operated by 11 TIF grantees participating in the evaluation, but we will look at two types of implementation of the TIF-funded intervention within each grantee's site: the programs with and without performance pay. This chapter provides details on what can be expected in the treatment schools (with performance pay) and under the counterfactual condition that the control schools are designed to represent (the same set of compensation-related reforms, but an unconditional bonus instead of performance pay).

Given that the programs we intend to study are still in their TIF grant planning year, this chapter cannot provide a comprehensive description of what will be in place at the time of the evaluation. Instead, we provide an overview of the elements required for each TIF grantee and examples of the range of the grantees' plans, as of November 2010, taken largely from grant applications. We want to emphasize that these programs are works in progress as districts and stakeholders refine their compensation reform packages. With these caveats, we list the 2010 TIF evaluation grantees in Table II.1, along with some program summary statistics drawn from the grant applications. Future reports on the impacts of performance pay under TIF will not associate findings with specific districts or grantees.

We begin this chapter by describing elements of TIF programs that will be common to treatment and control schools and conclude with a description of the performance-based pay offered only in treatment schools.

A. Context for the Evaluation: Common Elements of All TIF-funded Programs

The TIF grants were intended to provide grantees an opportunity to design and implement a comprehensive performance-based teacher support system to address student achievement, teacher and principal recruitment, and teacher and principal effectiveness in schools with at least half of students in poverty. The system is a departure from the status quo of education policy, and it is within this system that we seek to understand the impacts of performance pay. In other words, does offering performance pay lead to improved outcomes *relative to a set of teacher pay reforms that does not offer performance pay?*

Given that the counterfactual for the evaluation is not business as usual, it is useful to provide a description of the elements of TIF grants:

Table II.1. TIF Evaluation Grantees – 2010

Grantee	State	Program Name	Districts Involved	Schools in the Evaluation (by district)	Schools in the Evaluation (by grantee)
Arizona State University	Arizona	Ready for Rigor	Glendale Elementary School District Mesa Unified District	Glendale: 12 Mesa: 8	20
Chalkboard Project	Oregon	CLASS	Bend-LaPine School District Greater Albany Public School District	Bend-La Pine: 12 Albany: 8	20
Chicago Public Schools	Illinois	Chicago Public Schools Teacher Incentive Fund	Chicago Public Schools	16	16
Guilford County Schools	North Carolina	Mission Possible Expansion Program	Guilford County Schools	20	20
Iberville Parish Schools	Louisiana	BOOSTER	Iberville Parish Schools	10	10
Jefferson County Public School District	Colorado	Jeffco Strategic Compensation Pilot	Jefferson County Public School District	20	20
Michigan Association of Public School Academies	Michigan	TEAMS Project	Michigan Association of Public School Academies (various charter schools)	20	20
New York City Department of Education	New York	The Urban Excellence Initiative	New York City Public Schools	20	20
New York State Education Department	New York	NYS Teacher and Principal Performance-Based Compensation System	New York City Public Schools Rochester City School District Syracuse City School District Yonkers Public Schools	New York City: 20 Rochester: 20 Syracuse: 20 Yonkers: 8	68
Ohio Department of Education	Ohio	Ohio Teacher Incentive Fund	Cincinnati Public Schools	20	20
Winston-Salem/Forsyth County Schools	North Carolina	Project Star 3	Winston-Salem/Forsyth County Schools	16	16
Total			**15 districts**	**250 schools**	

Note: Participating districts and schools as of November 2010.

Required Elements

1. Performance-based incentives
2. Pay for additional roles and responsibilities
3. Targeted teacher professional development

Optional Elements (received competitive preference)

4. Incentives to retain and recruit effective teachers in high-need schools and hard-to-staff subjects

With the exception of the first element, which is the focus of the evaluation, all treatment and control schools in the evaluation will implement all required elements. We will describe performance-based pay later in this chapter. Below, we describe the other priorities and provide examples of how some grantees propose to meet them.

Not all schools funded by the 2010 TIF grant will participate in the study. The evaluation focuses on 11 grantees out of a total of 62 grantees funded in 2010. For each of the 11 evaluation grantees, we include 8 to 68 schools that meet the evaluation criteria (Table II.1) and no more than 20 schools per district per grantee. However, 3 grantees operate schools in more than one district participating in the evaluation.[4] One evaluation grantee is an association of charter schools in Michigan.

Although implementation of TIF varies from grantee to grantee, the program's essential and optional elements ensure that the compensation and career-ladder system in TIF schools differs markedly from what would be considered the status quo in public education. As grantees finalize their TIF programs, we will collect detailed data that will permit us to describe the context in which the evaluation takes place.

B. Required Elements for All Evaluation Schools

All evaluation schools must implement certain elements of the TIF program in order to be eligible for participation in the evaluation. Below, we describe the elements that will be common to all treatment and control schools in the evaluation.

1. Pay for Additional Roles and Responsibilities

Historically, teachers have been subject to a uniform salary schedule that determines salary based solely on years of teaching experience and education level. Teachers have earned extra pay outside regular contract hours only by coaching a sports team or teaching summer school. Such a structure provides little opportunity to compensate teachers who assume additional responsibilities and leadership roles during the regular school day, such as functioning as a master or mentor

[4] Two other grantees have several districts in their TIF program, but just one district in the evaluation.

teacher. TIF grant recipients were required to revise their pay structure such that grantees pay teachers for assuming roles such as mentor, master teacher, or tutor.

Illustrative Examples

Six grantees are using TIF funds to implement a TAP or similar model and establish a career ladder for teachers (Table II.2). Under such a program, teachers enter the profession as career teachers and then have the opportunity to advance to a mentor or master teacher. Mentor and master teachers are responsible for advising and observing other teachers in their schools. They often carry a reduced teaching load and sometimes receive full release time to focus on their leadership duties. Other grantees propose to follow a model similar to TAP in which teachers from within each school are hired to serve as teacher leaders responsible for leading regular teacher professional development seminars, observing other teachers, and providing opportunities for other teachers to observe their teaching.

Table II.2. Additional Roles and Responsibilities for Evaluation Teachers

Grantee	Additional Role Titles	Description of Additional Roles	Proposed Pay
Arizona State University	Mentor teacher, master teacher	Mentor new teachers, observe teachers, conduct professional development	$4,000 to $7,000
Chalkboard Project	TBD	TBD	TBD
Chicago Public Schools	Lead teacher, master teacher	Mentor new teachers, observe teachers, conduct professional development	TBD
Guilford County Schools	Value-added teacher leader, teaching standards teacher leader, model classroom teacher leader, mentor teacher	Assist teachers in understanding and implementing value-added data, coach teachers in understanding professional roles and responsibilities, host observations for new and struggling teachers, mentor new teachers	$1,000 to $2,000
Iberville Parish Schools	Mentor teacher, master teacher	Mentor new teachers, observe teachers, conduct professional development	$8,000 to $15,000
Jefferson County Public School District	Master teacher, mentor teacher	Conduct peer evaluations, provide general school leadership	TBD
Michigan Association of Public School Academies	Master teacher, mentor teacher	Mentor new teachers, observe teachers, develop professional learning communities	$2,000 to $4,000
New York City Department of Education	Master teacher, turnaround teacher	Mentor, coach, open classroom as "laboratory" for observation, develop lesson plans to be shared	15 to 30 percent of salary
New York State Education Department	Master teacher, leader teacher	Mentor new and ineffective teachers, provide video examples of effective instruction, develop professional development	12 to 14 percent of salary
Ohio Department of Education	TBD	TBD	TBD
Winston-Salem/Forsyth County Schools	Mentor teacher, master teacher, tutor	Mentor and support other teachers, tutor struggling students	$5,000

TBD – To be determined

All of the grantees that defined additional roles and responsibilities in their applications have created several leadership positions open to teachers. Teacher leaders receive extra compensation according to the level they attain, with the first step at $4,000 to $6,000 and the upper step at $7,000 to $15,000. Alternatively, some grantees plan to pay teacher leaders an increased fraction of their salary, such as a 12 percent pay increase for the first leadership step and a 14 percent increase for the second step. The maximum proposed value is 40 percent of salary.

Many applications do not indicate explicitly how the teachers who will receive pay for additional roles and responsibilities will be selected, although several limit eligibility to teachers rated as "highly effective." Information is also not yet available on whether teachers will retain the bonuses indefinitely or for only a limited term (such as a three-year appointment). We will obtain further information on these program features in our meetings with grantees and through a district survey that will be administered as part of the evaluation.

2. Teacher Professional Development

One of the major objectives of the TIF grant program is to increase educator effectiveness. In addition to financial incentives, grantees' compensation systems must provide "high-quality professional development targeted to needs identified through an evaluation system." Specifically, grantees were required to demonstrate that professional development provided through the TIF grant would be "directly linked to the specific measures of teacher and principal effectiveness included in the [compensation system]."[5] Grantees were advised that professional development should equip teachers and principals with the skills they need either to improve or maintain their effectiveness in raising student achievement. Likewise, grantees were required to demonstrate that they would offer teachers and principals professional development to help them better understand the measures to be used both as part of the compensation system and to improve student achievement. Finally, grantees must regularly assess the quality of their professional development and modify it as needed.

Illustrative Examples

Many evaluation grantees rely on the teachers and school leaders who receive compensation for additional roles and responsibilities to deliver professional development activities. Often, grantees plan to hold regular (e.g., weekly) professional development sessions in a small-group format (Table II.3). Grantees intend for these sessions to address current challenges faced by the teachers. In some cases, teachers will be expected to raise issues with the small group. In other cases, professional development will target the deficiencies identified during teacher observations or in students' formative assessments. Grantees also mentioned that teacher leaders will follow up with individual teachers and observe their practices. Several grantees' planned professional development sessions focus on orienting teachers to new data systems available for their use.

[5] The program rules appear in the *Federal Register*, May 21, 2010, p. 28742.

Table II.3. Professional Development for Evaluation Teachers

Grantee	Professional Development (PD) Resources	PD for Career Teachers Provided by	Hours/Frequency of Teacher PD	Training of PD Providers
Arizona State University	ASU will provide assistance to develop program	Master teachers	Weekly	None mentioned
Chalkboard Project	Partner with Oregon Dept. of Education's Direct Access to Achievement (DATA) project	TBD	TBD	None mentioned
Chicago Public Schools	TBD. Expect collaboration with an external provider	Unclear	Unclear	Targeted PD training for different teacher roles
Guilford County Schools	PD content is based on the NC Professional Teaching Standards, application of Education Value-Added Assessment System, and individual teacher value-added data	Master and lead teachers	Unclear	Master and lead teachers receive PD during planning year and during annual 3-day retreat
Iberville Parish Schools	Will use PD model developed by TAP	Mentor and master teachers and principals	Weekly	None mentioned
Jefferson County Public School District	None mentioned	Mentor and master teachers	Unclear	Master teachers and principals attend 5-day summer training on observation tool
Michigan Association of Pubic School Academies	PD will be offered through professional learning communities model	Mentor and master teachers	At least one hour/week	None mentioned
New York City Department of Education	Will use the ARIS Learn System to store professional development tools and create individualized developmental strategies	Master and turnaround teachers	Unclear	Executive and assistant executive principals will tutor new executive/assistant executive principals
New York State Education Department	None mentioned	Network teams	TBD	None mentioned
Ohio Department of Education	Might partner with Battelle for Kids	Unclear	Unclear	Will use focus groups consisting of highly effective teachers to identify best practices to share with other teachers
Winston-Salem/Forsyth County Schools	Learning Focused Model will be used. PD will also be available around a modified response-to-intervention model	District's Professional Learning Team	Unclear	Consulting agency will provide ongoing leadership training

C. Optional Element: Increased Recruitment and Retention of Effective Teachers in High- Need Schools and Hard- to- Staff Subjects

The TIF grant competition gave competitive preference to TIF designs that addressed the need to recruit and retain effective teachers in high-need subjects and hard-to-staff schools. As part of their grant applications, grantees had to provide evidence that they were targeting teachers who had demonstrated effectiveness (or were expected to be effective) and who were teaching in high-need schools and hard-to-staff subjects.

Illustrative Examples

Grantees are implementing TIF only in schools defined as high-need schools—those schools where at least 50 percent of the student body is eligible for free or reduced-price lunch. Therefore, many evaluation grantees argue that the entire TIF system is intended to provide incentives for teachers in high-need schools, with no additional incentives offered beyond the performance-based incentives and the pay for additional roles and responsibilities (Table II.4). Of those grantees planning to provide incentives for teachers working in high-needs subjects, few provided details on how they would ensure the effectiveness of newly hired teachers. Only one grantee explicitly addressed the issue of attracting more effective teachers by providing a 10 to 15 percent recruitment bonus to any new teacher who can demonstrate two consecutive years of exemplary student growth upon hire. Four grantees offer additional incentives for hard-to-staff subjects, with bonuses ranging from $2,000 to $10,000.

Although implementation of the TIF program varies from grantee to grantee, the program's essential and optional elements ensure that the compensation and career-ladder system in TIF schools differs markedly from what would be considered the status quo in public education. As grantees finalize their TIF programs, we will collect detailed data in order to describe the context in which the evaluation takes place.

D. Treatment Condition in the Evaluation

The evaluation design is an experiment in which we will randomly assign schools within a district to either a treatment or control condition. The treatment condition will implement performance-based incentives; the control condition, which will not implement differentiated performance incentives, will instead provide an across-the-board bonus of one percent of base pay. Grantees were informed that school status could not change during the course of the grant. In other words, schools offering performance-based compensation must continue to offer it for the life of the grant; schools that do not offer performance-based pay may not offer it at any time during the grant period.

The final notice for the TIF grant required that performance-based incentives must be:

1. Substantial in size, challenging to obtain, and differentiated
2. Based on an evaluation system that gives substantial weight to student growth
3. Based on an evaluation system that includes observation ratings linked to multiple observations per year

Table II.4. Incentives for Hard- to- Staff Subjects and High- Needs Schools

Grantee	Offering Separate Recruitment Incentive?	Positions Eligible for an Incentive	Proposed Recruitment Incentive Range
Arizona State University	Yes	Special education, secondary math and science	$5,000
Chalkboard Project	TBD		
Chicago Public Schools	Yes	Special education, math, science, bilingual education	$3,000
Guilford County Schools	Yes	Principals, special education, math, science, ESL, upper elementary, middle school language arts	10 to 15 percent of salary
Iberville Parish Schools	No		
Jefferson County Public School District	No		
Michigan Association of Public School Academies	Yes	TBD by local schools	$5,000
New York City Department of Education	No		
New York State Education Department	No[a]		
Ohio Department of Education	TBD		
Winston-Salem/Forsyth County Schools	Yes	New teachers in math, science, ESOL, exceptional children	$10,000

[a]No recruitment incentives are offered as part of the TIF grant, although some evaluation schools already offer recruitment bonuses.

Below, we describe each of these characteristics and provide illustrative examples of how some grantees have met them.

1. Substantial Compensation

If TIF funds are to influence teacher behavior, they must be sufficiently generous to provide incentives for greater effort or motivate new teachers to enter high-need schools or, ultimately, the profession. For example, if performance awards are small and paid to nearly all teachers, we would not expect substantial changes in teacher behavior. As a guideline, the grant notice used five percent of teacher and principal salaries as examples of average payouts that would be considered substantial for these positions. It also indicated that teachers and principals should have a reasonable expectation that they could earn substantially more than the average. As an example, some teachers and principals should expect that they will earn a bonus that is three times the average.

Along with being substantial in size, incentives must be relatively challenging to obtain but still attainable. In providing guidance on the types of criteria for payouts that would be deemed challenging, the grant notice gave an example of a criterion in which "payments are made only to [principals and teachers] who perform significantly better than the current average performance among study schools in the [local education agency]" (p. 28743).

Illustrative Examples

Evaluation grantees proposed a wide range of performance incentives for teachers and principals (Table II.5). Many districts set the maximum performance award as a percentage of base salary. When possible, we used the average salaries in these districts to calculate the maximum value for the average teacher so that we could use a common metric for all districts. The salary figures indicate a wide range in the amount that grantees have proposed for performance incentives. On the low end, one grantee proposed a maximum incentive of $3,000. In contrast, the proposed maximum was $20,000. Some of the variation may be muted during the planning year as grantees revise their programs.

Table II.5. Summary of Performance- Based Incentives

Grantee	Discrete or Continuous Performance Incentive	Maximum Value of Incentive
Arizona State University	Continuous	$3,000
Chalkboard Project	TBD	TBD
Chicago Public Schools	TBD	TBD
Guilford County Schools	Discrete steps	Up to $13,500 for teachers and $15,000 for principals
Iberville Parish Schools	Unclear	$6,000
Jefferson County Public School District	Discrete steps	$20,000
Michigan Association of Public School Academies	Continuous	15 percent of base salary
New York City Department of Education	TBD	TBD
New York State Education Department	Continuous	14 percent of base salary
Ohio Department of Education	TBD	$4,000
Winston-Salem/Forsyth County Schools	Discrete steps	$4,500 (teachers) and $5,000 (principals)

One difference among grantees is whether the performance-based compensation is discrete or continuous. For example, some grantees have proposed that some teachers will earn a bonus of $5,000 and that the remainder will receive nothing. Others have several steps that teachers and principals can attain based on their observed effectiveness. Others have proposed a continuous gradient that allows teachers to earn progressively more based on their performance without any discrete steps along the way.

2. Substantial Weight to Student Growth

The grant notice stated that performance-based compensation must "give significant weight to student [achievement] growth, based on objective data on student performance" (p. 28732). Growth is explicitly defined as the change in achievement over one or more years. While the definition is fairly straightforward for grades and subjects covered by annually administered state tests, it is more difficult for teachers in untested grades and subjects. Grantees therefore enjoy some flexibility in defining growth in those situations, although growth must be objective and comparable across schools.

Illustrative Examples

Most evaluation grantees factor into their performance-based pay the gains of students in the teacher's own class as well as growth in the achievement of other students in the school. One grantee also includes a measure of the gains of students of the other teachers on a teacher's team (Table II.6). For teachers in tested grades and subjects, the proposed weight on achievement from the teacher's own class ranges from 30 to 50 percent of the performance bonus while the weight on school achievement ranges from 5 to 15 percent. In untested grades and subjects, the only growth in student achievement that factors into a teacher's performance compensation is the growth of students in their own school, often replacing the own-class component for teachers of tested grades and subjects. For example, a grantee might assign a teacher in a tested grade a 35 percent weight for a teacher's own class and 15 percent weight for the growth of students in the school, but a teacher in an untested grade would receive 50 percent based on school achievement growth. In one case, 90 percent of the incentive for teachers in untested grades is based on school-level achievement growth.

The issue of including student growth measures for teachers in tested grades versus untested grades is one that most grantees have addressed by increasing the weight on school-level student achievement growth for teachers whose students cannot be directly assessed. However, one grantee has included for teachers in kindergarten and grade 1 a measure of how well their former students perform in grade 2.

In the absence of classroom measures of student achievement growth, student achievement ratings for principals' are based on school-level growth.

As mentioned, grantees have some flexibility in how they measure growth. However, the grant competition gave competitive preference to grantees proposing to use value-added models. Among grantees that specified in their applications how they intended to measure student growth, the most common approach was to use a value-added student growth model already developed by a third party such as the SAS® EVAAS® system, a value-added system developed by the SAS Institute.

Table II.6. Student Achievement Growth in Performance- Based Incentives

Grantee	Weight of Individual Student Growth in Performance Incentive (for tested grades)	Weight of School-Level Student Growth in Performance Incentives	How Student Growth Is Measured
Arizona State University	30 percent	20 percent	Unclear
Chalkboard Project	TBD	TBD	TBD
Chicago Public Schools	TBD	TBD	2- to 3-year average value-added (no other details)
Guilford County Schools	Unclear	Unclear	SAS® EVAAS® value-added system
Iberville Parish Schools	30 percent	20 percent (50 percent for untested grades)	Value-added model piloted in state
Jefferson County Public School District	TBD; 50 percent will be based on student growth; the individual/school mix is uncertain		Colorado Growth Model (no other details)
Michigan Association of Public School Academies	30 to 40 percent (depending on teacher type)	30 to 90 percent (depending on teacher type)	Scantron Performance Series (no other details)
New York City Department of Education	TBD	TBD	Teacher Data Initiative (value-added)
New York State Education Department	TBD; up to 40 percent is based on student growth; the individual/school mix is uncertain		TBD, but will be value-added
Ohio Department of Education	TBD; 50 percent will be based on student growth; the individual/school mix is uncertain		SAS® EVAAS® value-added system
Winston-Salem/Forsyth County Schools	50 percent	50 percent (38 percent on team growth and 12 percent on school growth)	SAS® EVAAS® value-added system

3. Observation Ratings from Multiple Observations

In addition to student achievement, performance incentives must be based in part on ratings from objective teacher observations. In their applications, grantees were instructed that they must "(1) [u]se an objective, evidence-based rubric aligned with professional teaching or leadership standards and the LEA's [local education agency's] coherent and integrated approach to strengthening the educator workforce; and (2) provide for observations of each teacher or principal at least twice during the school year by individuals (who may include peer reviewers) who are provided specialized training" (p. 28733). Beyond these instructions, grantees may exercise flexibility in determining who will conduct the observations, which observation protocols will be used, and how the observations will be related to the performance incentives.

Illustrative Examples

Many evaluation grantees plan to rely on the teachers who assumed additional roles and responsibilities to conduct classroom observations (Table II.7). These teachers, sometimes called school leaders, will receive training on an observation rubric before the start of the school year and must then demonstrate inter-rater reliability before conducting observations. Many grantees did not name the rubric that they intended to use but implied that they would identify an established rubric during the planning year.

Table II.7. Classroom Observations in Performance- Based Incentives

Grantee	Weight of Observation in Performance Incentive	Number of Times Observed	Who Conducts Observation
Arizona State University	30 to 48 percent (depending on teacher leader status)	Unclear	Master teachers, mentor teachers, principals
Chalkboard Project	TBD	TBD	TBD
Chicago Public Schools	TBD	TBD	Lead teachers, master teachers, school administrators, department chairs
Guilford County Schools	Unclear	Minimum of 4 times annually	Principals, assistant principals, master teachers, mentor teachers
Iberville Parish Schools	50 percent	4 to 6 times annually	Master teachers, mentor teachers, principals
Jefferson County Public School District	50 percent	6 times annually	Principals (2 per year) and master teachers (4 per year)
Michigan Association of Public School Academies	10 to 40 percent (depending on teacher type)	2 to 3 times annually	Graduate students from Michigan State University
New York City Department of Education	TBD	Minimum of 2 times annually	Principals, assistant principals, department chairs
New York State Education Department	Unclear	Unclear	"Trained evaluators" (no other description)
Ohio Department of Education	TBD	Minimum of 2 times annually	"Trained and credentialed educators"; no other description
Winston-Salem/Forsyth County Schools	Teachers must meet a minimum threshold score on their observations to be eligible for incentive pay; observations do not otherwise factor into the incentive amount	2 times annually	Trained administrators; no other description

The weight of teacher observation scores in overall performance-based compensation varies but can carry up to 50 percent of the weight for the total performance award. In some cases, the weight also varies within grantee; some grantees plan to place greater weight on classroom observations for teachers in untested grades and subjects. For instance, one grantee intends to weigh teacher observations as 15 percent of the incentive award for teachers in tested grades and subjects but 40 percent for teachers in untested grades.

4. Other Measures

In addition to student achievement growth and classroom observations, grantees have some flexibility in providing incentives for teachers and principals to meet other benchmarks. For example, grantees could provide performance incentives to principals for meeting target graduation rates or college enrollment rates.

Illustrative Examples

Most evaluation grantees did not propose to base teacher performance incentives on factors other than student achievement growth and teacher observations (Table II.8). However, one grantee planned to include peer evaluations (independent of classroom observations) as a determining factor in teachers' incentive pay. Another grantee will include the achievement growth of students who were in the teacher's class in the previous year. While such growth is technically a student achievement growth measure, it differs from the typical student achievement growth to be used by other grantees.

Principals' incentives more often included factors beyond student achievement growth. For instance, grantees proposed to include teachers' evaluations of their principal's leadership, the percentage of students enrolled in advanced placement classes, graduation rates, college enrollment rates, and the number of disciplinary incidents in the school. Two other grantees planned to provide principal incentives based on criteria that were yet to be defined but would be consistent with their schools' achievement of certain school improvement goals.

E. Summary

During the planning year, grantees had opportunities to develop the substance and implementation of their incentives. For example, many TIF grant applications did not explicitly address how performance incentives would be structured for teachers not in tested grades and subjects. Likewise, many grantees had not fully determined how to structure principal incentives. What is clear from the applications, however, is a wide variation in how grantees planned to implement their TIF projects. The range in the size of performance-based incentives is also notable. These differences might play an important role in understanding site-specific impacts. For this reason, we intend to collect detailed data on the implementation of the TIF program for all grantees participating in the evaluation.

Table II.8. Other Factors in Performance- Based Incentives

Grantee	Other Factors in Performance Incentive	Description of Factors
Arizona State University	Yes	Part of a principal's incentive is based on TAP fidelity. Peer/teacher evaluations are also factored into teacher and principal performance incentives (ranging from 2.5 to 20 percent, depending on teacher's role).
Chalkboard Project	TBD	The application suggests that performance incentives might be based on self-assessments and professional development.
Chicago Public Schools	Yes	Principal incentives will include teacher survey evaluations. Teacher incentives will include student engagement (attendance). The weight of the components is to be determined.
Guilford County Schools	No	
Iberville Parish Schools	Yes	Principal incentives will include measures such as graduation rates, but the measures are to be determined.
Jefferson County Public School District	No	
Michigan Association of Public School Academies	Yes	For some teachers, incentives include the growth in student achievement for the number of their previous years' students who meet standards in the current year. Principals' incentives include a measure of whether principals met certain school improvement goals (to be determined).
New York City Department of Education	TBD	The application leaves open the possibility that other factors will be included.
New York State Education Department	TBD	The application leaves open the possibility that other factors will be included.
Ohio Department of Education	Yes	Principal incentives will include measures such as student attendance, percent of students in advanced placement classes, and number of expulsions. Principal incentives will also be based on knowledge and skills as measured by the Ohio Standards for Principals.
Winston-Salem/Forsyth County Schools	No	

III. DATA AND SAMPLE

The evaluation will focus on a purposefully selected sample of approximately 250 schools in 9 states. While the evaluation grantees are known, the study team will determine the precise set of schools during the 2010–2011 school year as it works with the grantees to finalize the school selection and random assignment process. In this chapter, we describe the data collection plans, the process by which schools and teachers enter the study sample, and how we intend to allocate schools to the treatment and control groups.

A. Data

To answer the study's research questions and better describe the treatment and its counterfactual, we will collect data from several sources. We will request administrative data from TIF evaluation districts on their teachers and principals (to track mobility, obtain background characteristics, and collect information on educator effectiveness and incentive payouts) and their students (to measure achievement and link it to previous achievement, student background characteristics, and teacher/school/subject/grade assignment). We will also use survey data gathered from study school teachers, principals, and TIF district representatives.

1. Overview of Data Collection Activities

In Table III.1, we present the data collection schedule. We will survey all 2010 TIF districts (including evaluation and non-evaluation districts) three times—in fall 2011, 2012, and 2014. To obtain more detailed information on the experiences of evaluation districts, we will conduct in-depth telephone interviews with district representatives in spring 2012, 2013, and 2015. We will also survey principals and a sample of teachers each spring, and then, in the summer/fall, we will collect administrative records data. The current plan for the TIF evaluation calls for data collection in 2011–2012 and 2012–2013. The study design assumes that data collection will continue through 2014–2015.

Table III.1. Timing and Description of Data Collection

Type of Data Collection	Mode	Timing	Research Questions Addressed
Surveys of All 2010 TIF Districts	Paper	Fall 2011, 2012, 2014	Implementation
Evaluation District Interviews	Telephone	Spring 2012, 2013, 2015	Implementation
Teacher and Principal Surveys	Web	Spring 2012, 2013, 2014, 2015	Implementation, intermediate impacts, and mobility impacts
District Records Collection	Electronic records	Summer/fall 2011, 2012, 2013, 2014, 2015	Implementation, mobility, and test score impacts

Specifically, the instruments will collect the following types of information:

- **District survey.** We will use the data from a district survey, administered at three time points, to examine the association between impacts and key program features. Data from the first survey will be used to examine specific features of the incentive program and to understand approaches districts used to obtain buy-in and compromises they had to make. We will use information from the second survey to explore districts' experiences in the first year of program implementation and changes they had to make. Finally, data from the third survey will be used to describe districts' experiences since implementing the TIF program and ascertain their plans for sustaining the program.
- **Principal survey.** The principal survey will be used to assess hiring practices, classroom assignments, knowledge and perceptions of TIF in the study schools, and how all of these factors might have changed over time; the survey will also supplement administrative data to be obtained from district records. Moreover, the principal survey can provide important insight on principals' motivation for remaining in, leaving, or entering a study school.
- **Teacher survey.** The teacher survey will be used to assess knowledge and perceptions of performance pay and related issues in the study schools, to determine how these factors might have changed over time, and to supplement administrative data to be obtained from district records. The teacher survey can also provide important insight on teachers' motivation for remaining in, leaving, or entering a study school. We plan to survey teachers in grades 1, 4, and 7 for the study, as described below.
- **Principal and teacher administrative data.** These data will be used to estimate the impacts of performance pay on educator mobility and recruitment. The data will also allow us to examine the association between educator characteristics and student and educator outcomes, and to describe the principal and teacher samples. We plan to conduct a census of all schools in the study.
- **Student records data.** We will use existing state or district test score data to estimate the impact of performance pay on student achievement, the key outcome of interest. Information on students' demographic and socioeconomic characteristics and their achievement test scores prior to the study school year will be used to describe the students in the study and to develop more precise impact estimates. To the extent possible, we will use student-teacher linked data to estimate teachers' value-added scores to better understand mobility of high- and low-performing educators.
- **District interview.** The semi-structured questions in the district interviews will allow us to collect more in-depth information than that collected from the survey, and to probe for clarification if necessary. We will use this detailed information to more thoroughly understand the program context for each evaluation district, implementation strategy, and challenges. The district interview will also allow us to clarify the extent to which educators in the evaluation districts already face strong performance-based incentives from other policies.

2. Sampling

The study sample includes the TIF grantees that were selected through the 2010 evaluation competition, and will comprise as many schools as the evaluation districts are willing and able to

include, up to the maximum of 20 allowed. Thus, the study will not statistically sample grantees, districts, principals, schools, or students. We will sample teachers for the teacher survey as explained below.

The teacher survey will be administered to a representative sample of teachers in the study schools. We plan to survey two types of teachers—those in tested grade/subject combinations and those in untested grade/subject combinations. We will need adequate representation of both tested and nontested grades/subjects because each group of teachers faces different incentives and we need to measure the impacts of performance pay on each group separately. Those in nontested grades/subjects may be eligible for bonuses based on performance measures that they only indirectly affect, whereas teachers in tested grades/subjects have a more direct effect on these measures. Moreover, we will focus separately on elementary and middle schools. At the elementary level, the teachers are typically in nondepartmentalized settings, meaning that the classroom teacher is responsible for all subjects (including both math and reading), whereas in middle school the teacher is typically responsible for one subject (such as math *or* reading).

We plan to draw a census of all teachers who are responsible for math or reading instruction in all study schools in grades four and seven. By sampling teachers in this manner, we can limit the amount of heterogeneity that could be confounded with the treatment effect. For example, in several analyses, teacher survey data will be linked with test score data to examine impacts of performance pay on differential mobility of high- and low-performing teachers. If all of the sampled elementary teachers come from a particular grade (and likewise for middle school teachers), then the test scores used in this analysis will already be comparable between the treatment and control groups within each district. On the other hand, if we sample teachers from several grades, sample sizes of teachers per grade will be somewhat small, and there may be the possibility of grade imbalance (among either teachers or students) between the treatment and control groups, and we would rely on different r modeling approaches to address this potential imbalance. Overall, we believe it is preferable to focus on specific grades, limiting the generalizability of our findings, but also avoiding having to adjust for grade imbalances or risk confounding of grade level differences with treatment differences.

We anticipate an average of four teachers per elementary school, six per middle school (three each in math and reading), and eight per K-8 school (or schools with both elementary and middle school grades). This is a total of 1,300 teachers (4 x 150 elementary + 6 x 50 middle school + 8 x 50 K-8 schools).

For nontested grades/subjects, we will focus on first-grade teachers in elementary schools and seventh grade science teachers in middle schools. We selected the first grade because it has full-day classes and is less likely to have standardized testing than grades two and three. We selected science because it is a well-defined subject that is not routinely tested, but for which retaining certified teachers is an important policy goal.

We will randomly select two first-grade teachers from every elementary school, two seventh-grade science teachers from every middle school, and two of each type of teacher from every K-8 school. We anticipate this will result in a sample of 300 teachers from elementary schools, 100 from middle schools, and 200 from K-8 schools, for a total of 600 teachers.

The teacher survey data is more useful if we are able to follow the same teachers over time as well as refresh the sample with teachers who are new to the TIF schools, to learn about the types of

teachers each school is recruiting. Therefore, our sampling plan for the follow-up surveys is to survey 100 percent of leavers, 100 percent of replacement teachers, and approximately 75 to 82 percent of stayers. If we assume that 15 to 20 percent will leave the sample, and an equal number replace them, then we must reduce the sample of stayers by 18 to 25 percent, or in other words, survey 75 to 82 percent of the stayers. This is equivalent to following the 285 to 380 expected leavers, adding a roughly equal number of replacement teachers, and following 1,140 to 1,330 randomly chosen stayers.

In subsequent years, we will continue to survey all leavers and their replacements, since they are a smaller group and are of great policy interest, and follow a shrinking percentage of the stayers. If the exit rates after each year are higher than we have anticipated, we will consider altering the sampling rates by mobility status in order to achieve adequate sample sizes in each group.

3. Topics Covered

The data collection instruments discussed above will cover a range of topics that align with the research questions and logic model in Chapter I. They will help us understand how the performance pay programs are implemented and what impacts they have on the intermediate and final outcomes of the study.

Implementation. Careful measurement of program implementation permits both an understanding of what the treatment contrast represents and the interpretation of impact estimates. It also helps policymakers learn lessons from the TIF experience so they can anticipate challenges and avoid or overcome them in the future.

All of the data sources listed above, in addition to the study team's review of all available program documentation, will inform the implementation analysis. For example, we will use survey data on whether teachers who were eligible for performance pay understood that they were eligible (and whether those who were not eligible correctly understood their status). We will use administrative data on the size and timing of bonus payouts as confirmed by survey data. We will use descriptions from district interviews, program documents, and technical assistance team reports to understand how and when teachers were evaluated, informed of the results, and paid for their performance.

Key implementation milestones include not only the teachers' and principals' understanding of incentive program rules and compensation, but also the timing and completeness of performance measurement and targeted professional development. We will measure through surveys the duration and frequency of general professional development activities, teacher mentoring activities, and classroom observations in both treatment and control schools so that we can characterize how these activities were increased as a result of offering performance-based pay.

We will also develop some absolute measures of program implementation (not just treatment relative to control) that we will use to describe, from the grantees' perspective, what was implemented and when. We will use the district survey and the evaluation district interviews in conjunction with TIF program documents and records to describe the nature of the incentive programs that were offered. This includes a description of all of the components, including those that support incentives, such as communication and professional development efforts, as well as data infrastructure efforts. These district-level data collection efforts will also focus on challenges to implementation, such as stakeholder opposition to performance pay, requirements that districts

obtain formal approval from stakeholder groups, or anything unanticipated that districts volunteer in response to probes during the semi-structured interview.

Intermediate outcomes. Intermediate outcomes, such as teacher and principal attitudes and behaviors, are useful leading indicators of possible longer-term impacts. We will measure these constructs by using surveys administered to educators in the same fashion in both treatment and control schools. The types of intermediate outcomes include teacher reports on collaboration (to measure whether performance pay might undermine collaboration) and teacher self-reported satisfaction.

The most important intermediate outcome is the retention of teachers and principals. As discussed in Chapter I, final outcomes (student achievement, discussed below) are directly influenced by the mix of teachers who ultimately choose to teach in targeted schools. The data collection effort will inform the study of the rates at which teachers are retained and the characteristics of teachers who are recruited, retained, and lost to attrition in each year of the study.

Final outcomes. The final outcome of interest for the TIF evaluation is performance in the classroom measured in terms of growth in student test scores attributable to the intervention. The final outcome may be considered the total impact of the treatment contrast on student achievement. We posit (see Chapter I) that the impact is generated by a combination of two types of effects: composition effects (by changing the composition of the teaching workforce) and productivity effects (by encouraging teachers to improve their practice). Data limitations may prevent us from decomposing the total effect with great precision into separate productivity and composition effects, but—to shed light on the issue—we will examine teacher mobility and the types of teachers who move schools and leave the profession.

To measure the impact on student achievement, we will obtain test score data from each grantee at the end of each year of the evaluation. We will also obtain data on student background characteristics, such as eligibility for free or reduced-price lunch, special education status, and race/ethnicity, for use as covariates in the test score impact estimation.

B. Selection of Districts, Schools, and Teachers

Selection of districts and schools. Applicants to the TIF grant program selected the schools for the evaluation, and then the grantees were selected with input from a grant selection committee based on the extent to which their applications met the criteria set forth in the grant competition announcement. In Chapter II, we list the 11 evaluation grantees.

The program rules for TIF did, however, specify some parameters that helped shape the sample of schools for the evaluation. Schools had to include tested grades (elementary or middle school grades) and be "high need," which was defined as at least 50 percent of students eligible for free or reduced-price lunch. District superintendents and principals of potential study schools had to consent to be part of the grant application and to participate in the study, although schools will be added and replaced during the planning year (2010–2011) as needed. While not a grant requirement, some districts require school staff or union consent to participate; we will use the 2010–2011 planning year to obtain consent.

The selection process has implications for the study design. The study sample will not be representative of all districts in the nation but will instead consist of those that evidenced the need

for a teacher incentive program and undertook a successful grant-writing effort. Moreover, among all TIF grantees, the study sample includes only those that applied to, and were selected for, the national evaluation. Though not necessarily representative of all districts or all TIF grantees, the study sample may still be indicative of the type of school that might face the same policy choice tested by the study: teacher pay reforms with or without performance-based compensation.

Selection of teachers. The evaluation is interested in the impacts on all teachers who could have been influenced by incentives under the rules of the grantee's TIF-funded program. However, teachers are affected differentially. Teachers in tested grade-subject combinations may face a different performance measure and possibly different performance standards than those in nontested grade-subject combinations. Therefore, it is important to include teachers of both tested and nontested students.

Ideally, the study sample would comprise all teachers in the study schools, but the teachers represented in the data will vary by data source. The test score analysis will necessarily capture the impact on teachers in tested grades/subjects. The analysis of other outcomes will be based on surveys and administrative data. As discussed, we will survey a selected sample of teachers and a census of study school principals. The analysis of teacher and principal mobility will capture all teachers covered by the teacher and principal surveys, plus any educators captured in the administrative data provided by grantees. We have designed the teacher survey sample to provide a sufficient sample for estimating impacts in tested and nontested grades and subjects.

C. Random Assignment of Schools

Our overall approach to ensuring that inferences about performance pay are based on causal relationships is to use a random assignment process that guarantees that schools do not differ systematically on any dimension except the offer of performance pay. Below, we describe procedures for carrying out random assignment of schools in order to maximize the precision with which we can test hypotheses about the impacts of performance pay.

1. Overview of the Process

Stratification. We plan to make random assignment of schools within district and grade span (elementary or middle grades). We will then use school background characteristics to stratify the schools so that random assignment may be conducted within homogeneous blocks of schools, where feasible. We will define the blocks in such a way that we would expect to reduce within-block variation in latent outcomes as much as possible by increasing the between-block variation in latent outcomes. Latent outcomes are the outcomes that would be realized in the future in the absence of the intervention. Given that latent outcomes may not be observed, we propose to use school characteristics predictive of latent outcomes, such as baseline measures of the same outcomes (test scores and teacher retention) or easily measured characteristics such as student poverty, race/ethnicity, and school size.

We will also take into account factors that help the TIF program achieve important goals. For example, we may wish to stratify by school size so that the average size of the schools in the treatment group, and hence the size of the pool eligible for meaningfully large bonuses, does not depend on randomization. To illustrate why the size issue is important, we note that failure to stratify by size could result, by chance, in a treatment group that includes far more teachers than in the control group, requiring the TIF grantee to pay more in bonuses than it might have budgeted.

Another issue is the possibility that students move from control to treatment condition or vice versa when they graduate from elementary to middle school. Therefore, we also plan, where necessary and feasible, to group schools by feeder pattern so that an elementary school that feeds into a middle school will be assigned to the same treatment status as the middle school, if both are in the study.

Timing. We plan to conduct random assignment as close as possible to the start of program implementation. Informing schools of their randomization status too early might mean that interceding changes, such as school closures, principal turnover, or changing faculty preferences, would undermine a school's commitment to its original assignment, thereby potentially weakening or undermining the experimental design. On the other hand, informing schools of their assignment too late could interfere with program implementation. For example, if recruiting teachers or selecting and training classroom observers is part of the intervention and must take place in the spring before the first year covered by the incentives, then random assignment would have to be completed early enough to allow these activities to occur on schedule.

We plan to conduct random assignment between February and June 2011, discussing with each grantee the preferred timing so as to strike the right balance between early notification to aid implementation and late notification to minimize non-compliance or dropout. We may permit some exceptions to allow early random assignment for districts where principal turnover, teacher transfers, and school participation are carefully controlled and where implementation activities start earlier than February.

2. Protocol

Step-by-step procedure. Below, we list the step-by-step procedure we intend to follow to ensure a high-integrity design.

1. Within each evaluation district (or charter organization if not within the same district), we will first identify schools eligible for random assignment.
2. We will ensure that we have principals' consent and teacher buy-in, including district support, so that staff or leadership turnover does not compromise a school's commitment to carrying out its experimentally assigned status. Consent to participate in the study and abide by the random assignment is a precondition of random assignment.
3. We will obtain data on school characteristics to use for stratification: grade span (elementary/middle/high), , percent of students receiving free or reduced-price lunch, percent African American, percent Hispanic/Latino, total enrollment, feeder pattern, and, if available, data on average test scores and teacher turnover.
4. We will stratify according to the variables listed above. If any cells contain a singleton school, we will collapse the table by eliminating or collapsing a stratifier variable.
5. (Optional). It may be desirable to seek grantee input on stratification, thereby reducing the possibility that grantees will raise an issue after random assignment that could have been resolved before assignments were made.
6. Within strata, we will assign each school a random number and sort schools by that lottery number from lowest to highest.

7. If there are N schools in a cell and N is an even number, we will assign the first N/2 schools to the treatment group and the remainder to the control group. If N is an odd number, we will do the same but assign the "middle"-ranked school ((N+1)/2) according to the third digit of the random number. If that digit is odd, we will assign the school to the treatment group. If it is even, we will assign it to the control group.[6]

Protocol enforcement. The study team will measure the extent of program implementation in both treatment and control schools to determine whether it is consistent with the intended study design. For instance, we will monitor the types and timing of the information that teachers receive about the incentives for which they are eligible. We will also aim to track the mobility of students and teachers between schools of different study status—including treatment, control, and non-study schools—and the timing of their moves, in order to ascertain whether this mobility affects the interpretation of the impact estimates, as discussed below.

3. Design Caveats

The treatment and control schools will be operating within a common labor market, which can complicate the interpretation of the experimental results.[7] The mere presence of treatment schools may change the composition of the faculty in the control schools. For example, a high-performing teacher who might have opted to work in the control school could be lured away to a nearby treatment school with the expectation of earning a large performance bonus. Such a move on the teacher's part would not only improve outcomes for the treatment group but also worsen them for the control group, thereby overstating the impact we would have observed had the treatment and control schools been in separate labor markets.

Another consideration is the possibility that students will base school choice on treatment status. In this case, differences in student achievement could reflect both the impact of performance pay on teachers and the impact on the types of students who select a given school.

We will take several steps to help understand the degree to which teacher and student sorting might influence the interpretation of impact estimates. One is to track the mobility of teachers and students across schools. We will survey teachers who are new to the study schools to find out which schools they considered when making their decisions to transfer or apply to specific schools. We will also seek administrative data that allow us to determine whether a teacher moved from a treatment to a control school or vice versa. If we see teachers moving between treatment and control schools after the initial year, we can estimate what the program impact would have been if teachers were

[6] In districts with more than one odd-numbered cell, we will achieve a balanced sample by conditioning the assignment of the middle-ranked school as follows: we will pair the randomization blocks arbitrarily and, within each pair, assign one of the middle-ranked schools according to the third-digit rule and assign the other to the opposite condition as the first. In districts where we have an odd number of odd-numbered cells, we will first ask the grantee if it has a preference for the balance to favor treatment or control. In the case of a preference, we will follow that preference. If not, we will randomly determine whether the "extra" school will be assigned to treatment or control. One such mechanism is to use the random number itself. A random number (presumably drawn from a uniform [0,1] distribution) greater than 0.5 would be assigned to treatment and otherwise to control.

[7] Such interaction could violate the stable unit treatment value assumption (SUTVA), which posits that potential outcomes for any given study unit are unrelated to the treatment status of other study units. SUTVA is required for unbiased estimation of treatment-control differences under the Rubin Causal Model (Rubin 1974; Angrist et al. 1996).

classified by their initially assigned treatment status (Angrist et al. 1996). This sensitivity test can shed light on the role that teacher mobility plays in generating or attenuating a total impact.

Within each evaluation district, principal survey data will also provide information on the degree to which teachers are likely to choose their schools. Principals will be asked to report the amount of input and autonomy they have in teacher hiring, which will shed light on the extent of centralization in each district's system of teacher recruitment. The impact estimate will be easier to interpret in locations where teachers have less flexibility to comparison-shop for jobs across the entire district.

IV. DATA ANALYSIS

We will conduct several types of data analysis to address the research questions. First, impact analyses that exploit the random assignment design will yield unbiased estimates of the impacts of performance pay on student achievement and educator mobility. These experimentally-based estimates will also be key to an analysis of compositional changes in the teaching force triggered by performance pay. Second, correlational analyses will describe the association between performance pay features and impacts on key outcomes. Third, a set of implementation analyses will document treatment-control contrasts along several dimensions—including compensation expectations, actual payout distributions, and educator attitudes—relevant to identifying potential reasons for any observed impacts on student achievement. We will also tabulate descriptive statistics on implementation at the district level, examining both the group of 15 districts within the 11 evaluation grantees as well as the complete group of 186 districts within all 62 TIF grantees, including those not participating in the evaluation. Fourth, to complement the experimental analysis of the impacts of performance pay, we will tabulate outcomes for TIF program schools and compare them to outcomes for schools that are not part of the TIF program at all.

This chapter details our plans for each type of analysis. First, we describe several analyses that will document the implementation of TIF programs. Second, we specify the outcome variables on which we will conduct the experimental impact analyses. Next we describe the empirical methods for the primary impact analyses as well as detailed analyses of teacher mobility that build on the primary impact analyses. We then specify the smallest effects that can be reliably detected by these impact analyses. The chapter concludes with a discussion of the correlational and nonexperimental analyses.

A. Implementation Analyses

A diverse set of analyses will give rise to a detailed understanding of TIF programs, including the array of program features that the TIF grantees implement, the strategies employed and challenges encountered with implementation, and the demographic and policy environment in which the programs are implemented. This understanding is critical for interpreting the study's impact findings, placing them into the appropriate context, and identifying potential sources or reasons for any observed impacts. A systematic analysis of TIF program implementation can also yield lessons for the development and improvement of similar programs.

The analyses will draw upon multiple sources of data, as discussed in Chapter III, to examine several dimensions of program implementation. First, we will describe characteristics of TIF districts with data from the district survey and the Common Core of Data. Second, on the basis of responses to the district interviews, we will identify policies other than the TIF program that shape incentives faced by educators in the evaluation districts. Third, we will use survey responses from all TIF districts—supplemented with more in-depth information on evaluation districts from district interviews, administrative program records, and educator surveys—to summarize important components of TIF programs. Fourth, we will analyze implementation experiences on the basis of responses to the district surveys and interviews. Fifth, using educator survey data from the evaluation sites, we will document treatment-control differences in program implementation and other intermediate factors that could give rise to, or shape the extent of, resulting differences in the study's key outcomes. Next, we discuss each type of analysis.

Grantee district characteristics. A key purpose of describing grantee district characteristics is to understand the extent to which evaluation districts are representative of the full population of TIF districts. Accordingly, we will tabulate the average characteristics of three groups: (1) evaluation districts; (2) main districts; and (3) the combined population of all districts. The tabulations will document several institutional and demographic characteristics of the grantee districts, including size, urbanicity, racial and ethnic composition, and prevalence of free and reduced-price lunch receipt.

Policy context. To the extent that educators in the grantee districts already face strong performance-based incentives from other policies, there might be a smaller margin in which TIF-funded incentives could produce additional productivity gains (see Chapter I). Thus, within the evaluation districts, the policy environment might offer clues to explaining the presence or absence of impacts in the experimental study. We will draw upon responses from the district interviews to describe non-TIF incentive programs present in the evaluation districts. District staff will be asked to identify other programs in the evaluation districts—including federal, state, district, and privately-sponsored initiatives—that have introduced reforms of educator compensation similar to those in the TIF program.

Features of TIF programs. Describing the TIF programs with uniformly-defined measures of program characteristics is essential for identifying similarities and differences among the grantees' approaches to compensation reform. Although we discussed certain dimensions of the evaluation grantees' proposed programs in Chapter II, data from the district surveys and interviews will enable us to document comprehensively—within a uniform framework—the actual approaches to educator evaluation and award determination implemented by all TIF districts. Dimensions on which we expect to summarize and compare the TIF programs include the following:

- **Program coverage:** We will ascertain the grade spans and types of staff covered by the TIF programs, focusing especially on how performance measurement and compensation determination (discussed below) differ for teachers in tested and untested grade-subject combinations.
- **Evaluating educator effectiveness:** We will document typical ways in which educators are evaluated on the basis of student achievement and classroom observations. Given current policy interest in value-added measures, we will determine the prevalence with which TIF grantees use value-added measures as the achievement-based component of the performance measures. In addition, classroom observation measures used by TIF grantees will be characterized along various dimensions, including the length and frequency of observations and the types of staff who conduct the observations.
- **Determinants and structure of performance pay:** We will summarize the performance pay components of TIF programs in several ways. First, we will identify the types of performance measures—including student achievement at the classroom and school levels, classroom observations, peer reviews, and supervisors' ratings—that determine performance-based payouts. Second, we will ascertain the weight on each type of measure in calculating payouts; the weights can either be explicitly reported by the grantees or implied by the maximum payouts stemming from each measure. Third, we will summarize the distribution of expected payouts as reported by the grantee districts. Grantee districts will estimate the percentages of educators expected to receive each of

various ranges of performance-based payouts, and we will use these figures to calculate the degree of within-district dispersion in expected payouts.

- **Other TIF-funded compensation:** We will describe other factors that enable educators to receive additional compensation from the TIF program. For each TIF district, we will determine whether educators can receive additional pay for working in high-need schools, teaching hard-to-staff subjects, and attending professional development activities. Maximum award amounts attributable to each of those optional components—as well as components mandated by the TIF grant, such as additional pay for assuming extra roles and responsibilities—will gauge the relative emphasis on the component in determining compensation.
- **Targeted professional development:** Given that grantees are required to provide professional development (PD) for educators to understand their evaluation results and improve deficient areas identified by these evaluations, we will describe features of the TIF-funded PD. For all grantees, we will use district survey responses to document representative topics covered by the PD and the reported amounts of time allocated to PD. In addition, for the evaluation grantees, we will summarize principals' reports of the percentage of teachers receiving PD in each of various topics related to understanding and making use of effectiveness measures.

Implementation experiences. The district surveys and interviews provide a unique opportunity to draw lessons from numerous school systems' experiences in implementing compensation reform. We will extract common themes among the implementation challenges faced by the grantees and the strategies they used to address them. In addition, we will document variation in implementation experiences across grantees and identify patterns, such as institutional factors associated with differences in implementation. Our analyses will cover the following facets of TIF implementation:

- **Challenges in obtaining stakeholder support:** We will focus particularly on the challenges and constraints that the grantees face in securing the cooperation of principals and teachers. For instance, we will document whether each grantee must obtain formal approval from educators and/or their unions to carry out TIF programs. From the district interviews, we will further describe the stakeholders that are reported by the respondents to pose resistance to TIF implementation.
- **Strategies to obtain stakeholder support:** Potential channels for securing stakeholder support include disseminating program information, soliciting stakeholder input, and modifying the program design to address stakeholders' concerns. We will describe grantees' efforts on all of these dimensions. For instance, we will document the most prevalent modes of communication that grantees use to provide program information to school staff, unions, and parents. Moreover, for each grantee we will identify (1) the types of teachers' representatives (such as unions, non-union associations, or teacher committees) that provide formal input into the design or implementation of the program; (2) the venues through which the teachers' representatives provide such input; and (3) the stages of program development or implementation at which the grantee solicits this input. We will also document how common it is for grantees to modify their TIF programs to address concerns from various stakeholders, including principals, teachers at participating schools, and unions.

Treatment-control differences in intermediate outcomes. The central element of the study design is that one component of the TIF program—the presence or absence of performance pay—differs between treatment and control schools within the evaluation districts. To interpret any observed impacts on final outcomes, it is important to measure the actual treatment-control contrast in incentives stemming from the program differences. Similarly, examining treatment-control differences in various intermediate outcomes—such as educators' behaviors and attitudes, or the frequency of particular staff activities—can suggest potential pathways to catalyzing changes in student achievement and teacher mobility. Treatment-control differences in intermediate outcomes might result not only from educators' direct responses to performance pay, but also from differences in how any component of the TIF program is implemented in the two groups. In particular, because the stakes attached to performance evaluations are higher in the treatment group, district or program administrators might carry out various processes—such as classroom observations, communication of effectiveness ratings, and dissemination of program information—differently in treatment than control schools. Examining intermediate outcomes can suggest whether these implementation differences might be present.

We will use the educator surveys as the main source of data for these analyses. To determine whether educators in the treatment and control groups perceive different incentives—a precondition for impacts on final outcomes—we will document the extent to which educators in the two groups accurately identify their eligibility or ineligibility for the performance pay incentives. Similarly, treatment-control differences in additional compensation amounts that educators expect to receive will also capture the contrast in perceived incentives, and differences in actual amounts—from administrative program data—will be indicative of how the two groups will update their expectations for subsequent years. Beyond compensation differences, several other differences could be triggered by differential incentives or program implementation as intermediate steps toward affecting outcomes. For instance, we will examine whether teachers' attitudes toward their job—including overall job satisfaction, as well as perceptions of how the TIF program has influenced teacher collaboration and effort—differ between the study groups. We will also measure processes that might be implemented differently due to different stakes, including the amount and types of professional development, the ways in which classrooms are observed, and the frequency of teacher mentoring. From these analyses, we will be able to identify a diverse set of factors that can and cannot be ruled out as potential reasons for the observed impacts on the study's final outcomes.

B. Outcomes for the Impact Analysis

Student achievement, educator (teacher and principal) mobility, and characteristics of schools' teaching staff are the central outcomes on which we will estimate experimentally-based impacts of performance pay. These estimated impacts will serve as building blocks for estimating other impacts, including impacts on the composition of the teaching force as measured by teacher effectiveness. To ensure that the experimental impact estimates can be combined across grantee districts, the outcome measures themselves must be comparable across districts.

1. Student Achievement

A major challenge for the study is to standardize the measures of student achievement. As discussed in Chapter III, we will use students' scores on state or district assessments as measures of academic achievement. Assessments will generally differ and be measured using different scales across grade levels, subjects, and states; the TIF study includes both elementary and middle grades, examines math and reading/language arts achievement, and spans 9 states. In addition, special

populations, such as English language learners and students with disabilities, may take alternative tests or regular tests with accommodations, and students who repeat grades may take the same tests in successive years. All of these challenges must be considered when constructing the test score outcome variables.

To express test scores in the same units across all districts and grade levels, we will standardize test scores by creating z-scores within each combination of district and grade. That is, for each grade level, we will take the difference between a student's original test score and the districtwide mean score in that grade level, and divide the resulting difference by the statewide standard deviation of scores for that grade level. If statewide standard deviations are not available, we will use districtwide standard deviations.[8] Standardization will be done separately for each subject.

Expressing scores in common units allows impact estimates to be combined across grade levels and grantees and makes it possible to compare impact magnitudes between different grantee subgroups. Additionally, because standardization equalizes average outcomes across grade levels, it removes any spurious treatment-control differences in outcomes that might arise from chance differences in grade level distributions of the two groups.

In order for standardization to yield interpretable overall impacts, a standard deviation of student test scores in any one grade level within a district should represent the same increment of learning in every other district and grade level within the study. Because our sample spans several districts—from large, urban districts to small, rural ones—we prefer, whenever possible, to standardize using statewide rather than district-specific standard deviations.

Test scores of students who belong to various special populations, such as students who take regular tests with accommodations, will be standardized in the same manner as the scores of all other students. However, the impact analyses will control for a student-level covariate, if available, that indicates the use of test accommodations. Students who take alternative tests will not be included in any analysis because scores from alternative tests would need to be standardized from very different score distributions than those of regular tests, rendering the z-scores incomparable. Grade repeaters will be treated as if they have missing pretest scores because their prior-year test scores are not comparable to those of their grade-level peers. In analyses that require a non-missing pretest, grade repeaters will be omitted. They will be included, however, in analyses that do not use pretests as a covariate or that impute missing pretests. We will calculate treatment-control differences in the prevalence of alternative test-taking and grade repetition to ensure the main impact findings are not driven by these differences.

2. Educator Retention

Teacher and principal retention are outcomes of interest for two reasons. First, any turnover can be costly and disruptive. There are direct replacement costs associated with recruiting new staff, and there are indirect costs in adjusting to new procedures, getting to know students and their families, and adopting a new school culture. Second, turnover can have a net positive or negative impact on learning depending on whether those who leave are less or more effective than

[8] The use of districtwide or statewide means makes no difference to the impact estimates, given that indicators for randomization strata will be controlled for in the estimating equation, as described later.

replacement teachers or principals. This study will therefore measure retention rates overall as well as those of various types of teachers.

Because teacher mobility can take several different forms, it is important to construct outcome measures that reflect the types of mobility of greatest relevance to the hypothesized effects of performance pay. Each of the following types of departure represents different forms of teacher mobility: leaving one's initial school; leaving the group of schools within one's district that share the same compensation structure (that is, share the same treatment status in the study); leaving the district; and leaving the teaching profession. As discussed in Chapter I, a hypothesized mechanism by which performance pay can affect student outcomes is to make it more appealing for high-performing teachers to work in school environments with these incentives. On the basis of this hypothesis, performance pay is most likely to reduce high-performing teachers' rate of attrition from their schools and from the group of schools with the same compensation structure.

This study will focus primarily on retention of teachers in their initial schools. We will measure this using a dichotomous, teacher-level variable indicating whether a teacher stays within the same school from a specified initial year to a specified subsequent year. Retention of teachers at their initial school is more comparable across grantees than retention in treatment status groups, because some grantee sites have large districts with a few treatment schools and many schools not participating in the study, while others have just a few schools, all of which are participating in the study. Moreover, focusing on same-school retention facilitates comparisons with other evaluations that have examined this outcome (Glazerman et al. 2010). However, we will also conduct additional analyses of the other types of mobility: retention in a "team" (such as a grade level or department), in a study status grouping (the set of treatment or control schools), in the district, and in the teaching profession. These retention outcomes can be measured using binary variables (retained or not) or categorical variables for several mobility outcomes (e.g., indicating a switch into each of several school types within the same district, a different district, a charter school, or a private school).

Where possible, with each retention measure, we will examine the rates of retention overall and within categories of teachers as defined by their prior performance (value added) and professional background characteristics, as discussed below.

3. Characteristics of Schools' Teaching Staff

Performance pay has the potential to trigger changes in the *quality composition* of a school's teaching staff—that is, the mix of higher- and lower-performing teachers who choose to teach in the school. There are several ways of measuring the quality composition of a school's teaching staff. The most relevant compositional measure assigns to each teacher an *outcome-based* quality measure—capturing the teacher's pre-intervention effectiveness—and then aggregates this outcome-based measure across all teachers on a school's teaching staff. Obtaining impacts on this type of compositional measure involves combining several types of estimates; we discuss our approach to this analysis in section D below.

A complementary approach we will implement will examine impacts on the average background characteristics of a school's teaching staff. Constructing this type of outcome measure is straightforward: for each selected teacher characteristic, the outcome variable is a school-level average of the characteristic among the teachers in a given school. It is desirable to select measureable characteristics most likely associated with teacher effectiveness. In the large body of

literature on the association between teacher characteristics and student outcomes, teacher experience—especially in the first few years of teaching—is the single characteristic that appears to be consistently associated with teacher effectiveness (Hanushek and Rivkin 2006). Therefore, school-level averages of measures of teacher experience (such as average years of experience or percentage with more than three years) will be one variable on which we estimate impacts. We will also estimate impacts on other measures of average teacher characteristics, including those that capture teachers' certification status and the competitiveness of their undergraduate institution.

C. Estimation Methods for the Experimental Impact Analysis

1. Basic Estimation Method

Because of the study's experimental design, average differences in outcomes between the treatment and control schools are unbiased estimates of the impacts of performance pay. However, the precision of the estimates can be improved by using regression procedures to control for student, teacher, or school baseline characteristics that may explain some of the variation in outcomes not related to the treatment itself. In addition, the regression procedures appropriately account for the hierarchical structure of the data (when applicable)—in which outcome variables are measured for smaller units, such as students or teachers, nested within schools (the units of random assignment)—in carrying out hypothesis tests.

Because of the variation in performance pay features across grantee districts, we will use a flexible regression model that reflects the potential for impacts to differ across districts. Estimation of the model generates district-specific impacts that can then be aggregated to produce an overall estimate of the impact of performance pay. When outcomes are measured at the student or teacher level, we will estimate variations of the following model for the outcome y_{ijk} of student or teacher *i* in school *j* within district *k*:

$$(1) \quad y_{ijk} = \mathbf{R_{jk}}\boldsymbol{\alpha} + \beta_k T_{jk} + \mathbf{X_{ijk}}\boldsymbol{\delta} + \mathbf{Z_{jk}}\boldsymbol{\gamma} + u_{jk} + \varepsilon_{ijk}$$

where $\mathbf{R_{jk}}$ is a vector of indicators for combinations of grade levels and randomization strata; $\boldsymbol{\alpha}$ is a vector of grade-by-stratum fixed effects; T_{jk} is a treatment indicator; β_k is the impact of performance pay in district *k*; $\mathbf{X_{ijk}}$ is a vector of baseline individual characteristics with coefficient vector $\boldsymbol{\delta}$; $\mathbf{Z_{jk}}$ is a vector of baseline school-level characteristics with coefficient vector $\boldsymbol{\gamma}$; u_{jk} is a random school effect; and ε_{ijk} is a random individual error term. The district-specific impacts of performance pay, β_k, are the key coefficients of interest in equation (1). We will estimate equation (1) with ordinary least squares (OLS) using Huber-White ("sandwich") standard errors that account for school-level clustering.

Equation (1) will be adapted to suit each outcome variable. In particular, if the outcome variable is dichotomous, we will estimate the equation as a logistic regression. If the outcome variable is measured at the school level, then the regression equation will accordingly be estimated at this level (without individual-level covariates and error terms).

Our primary interest is in the overall, average impact of performance pay in the full study sample. There are several alternative approaches to calculating the average impact from the

regression estimates; each approach takes a weighted average of β_k across districts but differs with respect to the weighting scheme. If districts are weighted equally, then the average impact represents the expected impact for a randomly chosen district from the set of study districts; if districts are weighted by the number of schools or students, then the average impact represents the expected impact for a randomly chosen school or student, respectively, from the study sample. Because schools are the units assigned to the various compensation regimes in this study, we are interested in the expected impact for a randomly chosen school. Therefore, we will weight each district by the number of treatment and control schools when averaging the district-specific impacts.

2. Covariates

When estimating equation (1), we will control for several types of baseline characteristics—those that are measured prior to treatment assignment or that cannot be affected by treatment status—of study participants to enhance the precision of the estimates and control for any chance differences in observable characteristics between the treatment and control groups. Depending on the unit of analysis, the set of covariates can include student-, teacher-, and/or school-level variables.

In the analyses of student achievement, we will include student-level variables among the covariates in equation (1). While the variables to be included depend on the available data in administrative records, we expect the student-level covariates in the first year of the study will include prior-year test scores (*pretest scores*), gender, race and ethnicity, free and reduced-price lunch eligibility, special education status, English language learner status, grade repetition, and the use of test accommodations.

One issue we will consider is the entry grade—the first grade at which standardized tests are administered. Given that our main estimation sample consists of students in the third to eighth grades, pretest scores will be unavailable for third graders in any district without second-grade testing. It is currently unclear whether most of the study districts administer standardized tests in the second grade.

There are several alternative approaches to addressing this entry grade issue in estimation. (For discussion, we will assume that third grade is the entry grade). Three possible approaches are:

1. Pool third graders with all other grades in the same estimation sample but include an indicator for missing pretests—as well as an exhaustive set of interaction terms between this indicator and all other control variables—into the covariate set.
2. Estimate equation (1) separately for third graders (without controlling for pretest scores) and all other grades (controlling for pretest scores) and compute a weighted average of the impact estimates from the two samples, with weights based on precision or sample size.
3. Limit the main estimation sample to grades four to eight and conduct sensitivity analyses that allow for the inclusion of third graders.

We intend to conduct the analyses using all possible sample members and all available data. However, for presentation purposes, we will consider each of these approaches after assessing the available data. For example, if a large percentage of the potential study sample is in an entry grade, we will recommend the first option. If it is a smaller percentage, then we will recommend the third option. The second option would be a compromise.

When teachers are the unit of analysis—in particular, in the analyses of teacher retention—we will control for a number of teacher background characteristics. Teacher characteristics measured by the teacher survey or potentially available from administrative data include gender, race, certification status, route to certification, highest educational degree, level of teaching experience, and competitiveness of the teacher's undergraduate institution.

Regardless of the level—student, teacher, or school—at which the outcome is measured, school-level control variables represent important components of the covariate set. These variables will consist mostly of school-level aggregates of student- and teacher-level characteristics calculated from the students and teachers, respectively, present in the study schools at baseline. Among these school-level variables, the most important one for improving the precision of estimated impacts on student achievement is a school's average test score in the pre-intervention year (Bloom et al. 2007). Indeed, because unexplained variation in outcomes across schools is a key source of imprecision in school-level random assignment designs, controlling for school-level baseline variables is expected to account for most of the precision gains from covariate adjustment in estimating equation (1).

3. Impact Estimates in Later Study Years

Study schools are expected to remain in their treatment status groups for four years, which provides an opportunity for multiple years of analysis. Accordingly, we will estimate impacts of performance pay in each of the first four years of TIF implementation (referred to as years 1, 2, 3, and 4). We choose to examine each period separately because, for various reasons, impacts in later years might be larger than those in earlier years. For example, changes in educator performance and the composition of the teaching staff at treatment schools may be more pronounced after educators observe performance-based payouts from earlier years. Also, if cumulative impacts are the focus (as described below), they could grow over time as students in treatment schools experience longer total exposure to the treatment.

Our basic approach to estimating impacts will be similar across the four study years. We will estimate equation (1) separately for each study year. For most outcomes, including student test scores and the average characteristics of the teaching staff, the estimation sample in each year will consist of the cross section of students or teachers present in the study schools during the specified year. However, retention outcomes will be defined longitudinally. In particular, among teachers observed in year 1, retention analyses based on available administrative data will examine impacts on whether teachers return to their year 1 schools at the beginning of years 2, 3, 4, and 5; retention analyses based on survey data will examine impacts on whether these teachers are observed in their year 1 schools in the spring terms of years 2, 3, and 4. The advantage of using administrative data, rather than survey data, is the broader coverage, allowing us to include teachers in grades and subjects not selected for the study sample.

One difference between the analyses in year 1 and those in later years is that some covariates that are exogenous in year 1 are potentially endogenous in year 2 and beyond. Student-level pretest scores are the key covariates to which this issue pertains. If the treatment were to have an impact on test scores in year 1, then the inclusion of pretest scores in the year 2 analysis would cause this covariate to absorb part of the cumulative impact of performance pay. Thus, in the presence of pretest controls, impact estimates in year 2 (and beyond) would represent annual impacts of performance pay on single-year test score gains; impact estimates without pretest controls would represent cumulative impacts.

For several reasons, our main analyses in year 2 and beyond will exclude student-level pretest covariates in order to estimate the cumulative impacts of performance pay in an unbiased manner. As stated earlier, given that various cohorts within the treatment schools will experience multiple years of exposure to performance pay, it is of interest to determine whether any benefits from this exposure accumulate over time. Moreover, cumulative impacts are estimated from purely experimental variation in treatment status, given that the exogenous covariates are uncorrelated with—and thus do not account for—any variation in treatment status in expectation. In contrast, because an endogenous pretest covariate could be correlated with treatment status in expectation, choices for how to model the functional form of the pretest covariate could influence the estimated annual impacts on single-year gains.

Because estimates of cumulative impacts in year 2 and beyond will not control for pretest scores, they are likely to have lower precision than estimates in year 1 that control for pretest scores; however, the precision loss is not expected to be large as long as the school-level average of pre-intervention test scores is always included in the covariate set. Indeed, in analyzing the precision of school-randomized designs in five large urban districts, Bloom et al. (2007) found that controlling for school-level averages of pre-intervention test scores alone—even lagged by two or three years—yielded precision gains nearly as large as those achievable by controlling for students' own prior-year test scores.[9] As discussed in section E, the study's minimum detectable effects are expected to rise modestly when the student-level pretest is excluded. More recent research (Deke et al. 2010) suggests that the gains from this approach may be more modest, so we will reassess the strategy when we have more information about the schools participating in the study.

Although focusing on cumulative impacts can bring the advantages described above, it is important to note that they represent cumulative effects on schools—not students. For example, the cumulative impact in year 3 represents the effect of assigning a school to implement performance pay for three years on the school's average level of student achievement; within the school, there will be a mix of students who have been exposed to the treatment for 1, 2, or 3 years. Moreover, these cumulative impacts can reflect compositional changes in the student populations of treatment controls relative to control schools arising, for instance, from endogenous student mobility. Although this appears unlikely, we will estimate treatment-control differences in average student characteristics in each year of the study to assess whether such mobility might have occurred.

In contrast, the annual impact on single-year gains pertains to a well-defined interval of exposure—one year—for all students, and is thus informative of how much a student would benefit in one year from being in a treatment school during a particular year of the study. Because annual impacts on single-year gains may still be of considerable policy interest, we will provide supplementary estimates of these impacts.

4. Subgroup Analysis

In addition to estimating the average impact of performance pay in the full study sample, we will also conduct impact estimates for key subgroups of individuals. These subgroup analyses will

[9] For example, for study designs that randomize 60 elementary schools, Bloom et al. (2007) found that controlling for a twice-lagged school-level average of test scores yielded a minimum detectable effect size of 0.22 to 0.23; controlling for once-lagged student-level test scores yielded a minimum detectable effect size of 0.21 to 0.24.

help determine whether performance pay has particularly strong or weak effects on certain populations.

For analyses of student achievement, grade spans will define the primary set of subgroups in which we will estimate impacts. Specifically, we will estimate an average impact on students in grades three through five (elementary grades) and a separate, average impact on students in grades six through eight (middle grades).[10] These analyses by grade span allow for the possibility that the responsiveness of student achievement to changes in teacher quality might differ by student age. Indeed, prior nonexperimental evidence has found that variation in teacher quality explains a greater proportion of student achievement variance among elementary school students than among middle school students (Kane et al. 2008).

For analyses of teacher retention and other teacher-level outcomes, we will identify teacher subgroups based on characteristics that potentially shape teachers' incentives and their responsiveness to those incentives. In particular, we will estimate impacts separately for teachers in tested and nontested grade-subject combinations. These two groups are typically evaluated by different performance measures—and, thus, are subject to different incentives—given that only teachers in tested grades and subjects can be evaluated on the basis of their own students' test score performance. Moreover, we will also estimate impacts separately for early-, mid-, and late-career teachers. Because retention is especially low in the first few years of teaching (Ingersoll 2003), there is a greater potential for treatment-induced retention increases among early-career teachers. Finally, we will estimate retention impacts by measures of teacher performance or performance-related professional background characteristics. This will capture whether treatment improves the quality of teachers through a retention effect. Another approach to examining compositional effects is discussed next.

D. Estimating Compositional Changes in the Teaching Force

Given that any observed impacts of performance pay on student achievement could be mediated through the two channels—composition and productivity effects—as discussed in Chapter I, discerning the contribution of one channel is sufficient to gauge the contribution of the other.[11] Our approach will be to undertake a detailed analysis of the composition effects of performance pay. We focus on composition effects because an analysis of this channel can reveal several types of changes, including changes in the mix of teachers who enter, stay in, or leave the study schools. This will provide a rich picture of the range of responses to performance pay.

Although impacts on the quality composition of a school's teaching staff could be reflected in various observable changes, including changes in average teacher background characteristics, the most direct evidence would consist of changes in the mix of higher- and lower-quality teachers as gauged by outcome-based measures. However, the use of outcome-based measures to examine composition effects faces important data limitations. Ideally, these measures should only capture teachers' effectiveness in a pre-intervention policy environment so that they do not reflect any

[10] In schools that span multiple grade spans, students in the elementary or middle grades will be assigned to the elementary or middle grade subgroups, respectively.

[11] This is true as long as the two channels contribute additively to the overall impact.

productivity effects from the intervention. Because effectiveness measures are typically based on student test score gains, constructing such measures for the pre-intervention period would require two years of pre-intervention test score data. The number of districts that can provide this data is uncertain. Moreover, for certain groups of teachers, including those who enter the study districts during the intervention period, pre-intervention effectiveness is entirely unobservable. Therefore, it is advantageous to use an approach to estimating composition effects that can rely solely on effectiveness measures from the intervention period. We present such an approach next.

1. Conceptual Framework

A simple conceptual framework can highlight different types of teacher mobility that contribute to treatment-induced compositional change. Although, for simplicity, we focus on compositional change that occurs between years 1 and 2, the framework applies to any two years of the study.

For any school—regardless of treatment status—we consider three types of teachers: (1) those who leave the school at the end of year 1, referred to as *end-of-1 leavers*; (2) those who enter the school at the beginning of year 2, referred to as *start-of-2 entrants*; and (3) those who are present in the school in both years 1 and 2, referred to as *stayers*. Let $L_t^{(1)}$, $E_t^{(2)}$, and S_t denote the average effectiveness that end-of-1 leavers, start-of-2 entrants, and stayers, respectively, actually exhibit in the year *t* policy environment of their school or would have exhibited in that environment if their effectiveness could have been observed. Within the specified school, the extent of compositional change between years 1 and 2 is

$$(2) \quad \delta = (1 - R)(E_0^{(2)} - L_0^{(1)}),$$

where *R* is the retention rate between years 1 and 2. In other words, compositional change occurs within any school whenever end-of-1 leavers are replaced by start-of-2 entrants with a different average quality. Moreover, differences in quality between entrants and leavers change the average quality mix of the school's teaching force to a greater extent when turnover is more prevalent (that is, when *R* is lower). Because a teacher's quality should ideally be gauged by his or her effectiveness in a pre-intervention policy environment, we subscript $E_0^{(2)}$ and $L_0^{(1)}$ with t=0. For reasons related to the feasibility of estimation, described below, we rewrite equation (2) as

$$(2') \quad \delta = (1 - R)[(E_0^{(2)} - S_0) + (S_0 - L_0^{(1)})]$$

by subtracting and adding the average quality of stayers, S_0, within the brackets.

Given that δ captures the compositional change in any given school, the objective is to estimate the *impact* of performance pay on δ—that is, the average difference in δ between treatment and control schools. As seen in equation (2'), an impact on δ can result from any of the following types of constituent impacts:

- **Impact on retention:** Impact on *R,* the overall retention rate
- **Impact on the entrant-stayer difference in quality:** Impact on $(E_0^{(2)} - S_0)$, the pre-intervention effectiveness gap between start-of-2 entrants and stayers

- **Impact on the stayer-leaver difference in quality:** Impact on $(S_0 - L_0^{(1)})$, the pre-intervention effectiveness gap between stayers and end-of-1 leavers

These three constituent impacts highlight that performance pay could generate positive composition effects in a number of possible ways. It could lead to the recruitment of higher-quality entrants, which would be reflected in a positive impact on the entrant-stayer difference in quality; it could raise the likelihood that the least effective teachers leave or the most effective teachers stay, which would be reflected in a positive impact on the stayer-leaver difference in quality. Raising the retention rate could generate a positive, null, or negative composition effect, depending on whether entrants were of worse, equal, or better quality, respectively, compared to leavers.

Importantly, two of the constituent impacts—the impacts on entrant-stayer and stayer-leaver quality differences—should not be interpreted in isolation because they are interrelated. In particular, if performance pay induces schools to selectively retain more of their higher-quality teachers—reflected in a positive impact on $(S_0 - L_0^{(1)})$—then the rise in stayer quality would be reflected in a negative impact on $(E_0^{(2)} - S_0)$, even in the absence of any recruitment effect. Only if there is no impact on the stayer-leaver quality difference can the impact on the entrant-stayer quality difference be interpreted as purely arising from a recruitment effect.

Despite these limitations in interpretation, the central advantage of examining impacts on quality *differences* between mobility groups is to enable feasible estimation of the overall composition effect. Below we describe our approach to estimating each of the three constituent impacts in the overall composition effect.

2. Estimation Methods

Under the preceding conceptual framework, the following key assumption allows composition effects to be feasibly estimated: in a given school and year, productivity effects are assumed to be constant across all teachers in tested grades and subjects. For instance, within a specified control group school, we will assume that all such teachers in year 2 experience the same increment in effectiveness—relative to the effectiveness they would have had under no reforms—as a result of the control condition's package of reforms in year 2; an analogous assumption applies to the treatment group teachers. This assumption is not especially restrictive, as productivity effects might still differ across years and across schools—but simply must not differ within the same school and year. A critical implication of this assumption is that for any two groups of teachers in a particular school, the group difference in effectiveness observed during an intervention year is identical to the group difference that would have been observed if reforms had not been in place. In particular, unobservable pre-intervention differences in effectiveness between teacher mobility groups—the types of differences that we aim to estimate—can be feasibly measured from observed differences in the intervention period. Using this principle, we discuss next how variants of the basic regression model in equation (1) can generate estimates of the three types of impacts that contribute to the composition effect.

Impact on retention. We will directly estimate the impact on retention in the teacher-level impact analyses when we estimate equation (1) with a dichotomous retention indicator as the dependent variable.

Impact on the entrant-stayer difference in quality. The challenge of estimating impacts on the quality of entrants alone, $E_0^{(2)}$, is that the effectiveness of start-of-2 entrants can only be measured beginning in year 2, when entrants into treatment schools are already subject to performance pay incentives and might therefore already exhibit productivity responses to treatment. However, as long as stayers are assumed to exhibit the same productivity response in year 2 as entrants in the same school, then the effectiveness *difference* in year 2 between entrants and stayers, $(E_2^{(2)} - S_2)$, is equivalent to the effectiveness difference that would have been observed in a year 0 policy environment; that is, $(E_2^{(2)} - S_2) = (E_0^{(2)} - S_0)$. Using this assumption, our approach will be to estimate impacts on $(E_2^{(2)} - S_2)$, a fully observable quantity, as a proxy for impacts on $(E_0^{(2)} - S_0)$. To do so, we estimate the following variant of equation (1) when year 2 test scores are the outcome:

$$(3) \quad y_{ijk} = \mathbf{R_{jk}}\boldsymbol{\alpha} + \beta_k T_{jk} + \rho_k D_{ijk} + \lambda_k (T_{jk} \times D_{ijk}) + \mathbf{X_{ijk}}\boldsymbol{\delta} + \mathbf{Z_{jk}}\boldsymbol{\gamma} + u_{jk} + \varepsilon_{ijk},$$

where D_{ijk} is a dummy variable for whether student *i* is taught by a start-of-2 entrant (rather than by a stayer). The estimated coefficient, $\hat{\lambda}_k$, on the district-specific interaction term, $(T_{jk} \times D_{ijk})$, captures the extent to which, within district *k*, the entrants' effectiveness relative to that of stayers is more positive in the treatment group than in the control group. A weighted average of $\hat{\lambda}_k$ across districts provides an estimate for the average impact on the entrant-stayer difference in quality.[1]

Impact on the stayer-leaver difference in quality. Our approach to estimating the impact on $(S_0 - L_0^{(1)})$ parallels our approach to estimating the impact on $(E_0^{(2)} - S_0)$. Because year 1 is the first year in which effectiveness is likely to be observable for any group of teachers, we cannot directly observe either S_0 or $L_0^{(1)}$. However, assuming that stayers and leavers exhibit the same productivity effects in year 1, we will estimate the impact on $(S_1 - L_1^{(1)})$, the effectiveness gap in year 1 between stayers and end-of-1 leavers, as a proxy for the impact on $(S_0 - L_0^{(1)})$. As before, we will estimate equation (3), with the exception that the outcome variable now consists of year 1 scores, and D_{ijk} now denotes a dummy variable for being taught by a stayer (rather than by an end-of-1 leaver).

Scaling and combining the constituent effects. We will scale—that is, multiply—each of the three constituent impact estimates by an appropriate factor and sum them into an overall composition effect. As seen in equation (2'), any impact on $(1 - R)$ should be scaled by the prevailing entrant-leaver quality gap in the control group; likewise, any impact on $(E_0^{(2)} - S_0)$ or $(S_0 - L_0^{(1)})$ should be scaled by the prevailing nonretention rate in the control group.

[1] Because we are assuming that productivity effects are constant only within schools, we will also replace the district-specific treatment indicators with school dummies in equation (3) to account for heterogeneity in productivity effects across schools.

E. Minimum Detectable Impacts

Given the planned impact analyses, the study must include a sufficient number of schools in the treatment and control groups in order to be able to detect policy-relevant effects of performance pay on the study's key outcomes. Determining the appropriate sample size entails two key steps. First, we select the study's minimum detectable impact (MDI), representing the smallest true impact of performance pay that the study should be able to detect with high probability. Second, we calculate the number of study schools required to ensure sufficient precision for detecting the chosen MDI.

1. Selection of Minimum Detectable Impact

To select an MDI, we identify the smallest impact of performance pay that could be policy-relevant. Because the study examines several outcomes, we first focus on student achievement, the outcome most relevant to the study's first research question from Chapter I; we then assess the size of impacts on other outcomes that can be detected with the resulting sample.[13]

We use two benchmarks to select an MDI; one is based on the normal year-to-year growth in student achievement, and the other is based on the typical degree of variation in teacher effectiveness within schools. That is, we determine the smallest impact, expressed in standard deviations (SD) of student scores, representing both a meaningful proportion of annual student growth and a meaningful increment within the distribution of teacher effectiveness.

The first benchmark is based on evidence compiled by Bloom et al. (2008), who use norming data from seven nationally standardized tests to calculate the average spring-to-spring gain in student achievement between consecutive grade levels. Annual growth varies by subject and declines considerably with grade; on average, from grades three to eight in reading and math, student achievement rises approximately 0.37 SD in one year.

The second benchmark uses observed variation in teacher effectiveness within schools as a basis for gauging the practical significance of impact magnitudes. Compiling published estimates from several studies, Hanushek and Rivkin (2010) find that, on average, a student's test score rises by 0.16 student-level SD if he or she switches to a teacher in the same school who is more effective by one standard deviation in the distribution of teacher effectiveness. From these estimates, an impact magnitude can be interpreted as raising the effectiveness of the average treatment school teacher from the 50th percentile, for instance, to a specified new percentile within the teacher effectiveness distribution of a typical control school.

On the basis of these benchmarks, ED determined that the study should seek to detect, at a minimum, an impact of 0.09 SD on test scores. This MDI represents about one-quarter of a year (that is, about three months) of achievement growth on nationally standardized tests. Moreover, such an impact is equivalent to a shift in the effectiveness of the average treatment school teacher from the median to the 71st percentile of teacher effectiveness within a typical control school. While this threshold may be high, the cumulative impact over multiple years has a greater likelihood of reaching this magnitude than the impact in the first year alone.

[13] Because student test scores are already converted to standard deviation units, the MDI is equivalent to the minimum detectable effect size (MDES).

2. Required Sample Size

To detect an impact of 0.09 SD on test scores, the study requires a sample of 250 schools, split evenly between the treatment and control groups (Table IV.1). Within each treatment status group, we expect that approximately three-fifths of the schools will be elementary schools, one-fifth will be middle schools, and one-fifth will be K–8 or K–12 schools. With this sample size, if the true effect of performance pay on student achievement is 0.09 SD, the study has an 80 percent probability of rejecting the null hypothesis of zero impact using a two-tailed test at a five percent significance level.

For various subgroups of interest, the minimum impacts that the study can detect are somewhat larger but still within the range of feasible effects. Because the schools with elementary grades—including elementary schools and K–8 or K–12 schools—are expected to constitute about 80 percent of the school sample, the MDI for elementary grades is barely higher—at 0.10 SD—than the MDI for the full sample (Table IV.1). Fewer schools contain middle grades, and therefore the study will be able to detect impacts of 0.13 SD or higher in these grades.

Table IV.1. Minimum Detectable Impacts on Student Test Scores

Type of Sample	Number of Schools	Number of Students	Minimum Detectable Impact (in Standard Deviations of Student Test Scores)
Full Sample	250	89,000	0.09
Grades 3 to 5	200	44,000	0.10
Grades 6 to 8	100	45,000	0.13

Note: The calculations are based on the following assumptions: 80 percent power and a 5 percent significance level for a two-tailed test; 60 percent of schools in the sample will be elementary schools, 20 percent will be middle schools, and 20 percent will be K-8 or K-12 schools; each elementary school will contain 240 students in tested grades, each middle school will contain 740 students, and each K-8 or K-12 school will contain 320 students in tested grades; end-of-year test scores will be missing for 15 percent of students in tested grades; 13 percent of the total variance of student test scores will be between schools; covariates will explain 65 percent of the test score variance between middle schools, 40 percent within middle schools, 50 percent between elementary schools, 33 percent within elementary schools, 55 percent between K-8 or K-12 schools, and 35 percent within K-8 or K-12 schools.

3. Sensitivity of Precision Calculations

The MDI that can be realized by a sample size of 250 schools depends on several parameters that will be unknown until data are collected. Our expectations for the likely MDIs, as shown in Table IV.1, are based on an informed assessment grounded in actual parameter values from other recent large-scale evaluations. Nevertheless, because there remains the possibility that the realized parameter values in this study will differ from those observed in the past, we explore a range of reasonable values for key parameters and examine the MDIs that would be achievable under these alternative scenarios. In particular, we assess the sensitivity of the MDIs to the following parameters:

1. School-level and student-level R-square: the proportions of test score variance between and within schools, respectively, that can be explained by covariates
2. Intracluster correlation (ICC): the proportion of total test score variance that is observed between schools

3. Proportions of students and schools with missing data

It is particularly important to gauge the sensitivity of the MDI to the school-level R-square because this parameter has varied considerably across prior randomized trials (Deke et al. 2010) and, in this study, is expected to differ between year 1 (when student-level pretest scores are controlled for) and later years (when student-level pretests are excluded from the main analyses). With a school-level R-square of 0.4—a value somewhat lower than what might be expected when school-level average baseline scores, but not student-level pretest scores, are controlled for—the MDI is 0.10 SD[14] (Table IV.2). Thus, the MDI in year 2 and beyond is not expected to differ considerably from that in year 1. Even with a worst-case R-square of 0.2, the MDI would still be 0.12 SD, or about one-third of a year of average student achievement growth.

For plausible ranges of other key parameters, the MDI does not rise above 0.11 SD. Raising the ICC from 0.13 (the benchmark assumption) to 0.20 would push the MDI from 0.09 SD to 0.11 SD. For a fixed number of schools with available data, the MDI is largely insensitive to reasonable deviations from our benchmark assumptions regarding the proportion of students with missing test scores. However, the MDI would change to a greater extent if entire schools failed to supply the data needed for the study. If, for whatever reason, data were completely unavailable for 40 percent of the schools in the sample, then the MDI would be 0.11 SD.

4. Minimum Detectable Impacts on Other Outcomes

With a sample size of 250 schools, from which 1,900 teachers will be sampled for the teacher survey, the study will be able to detect policy-relevant impacts of performance pay on teacher retention. The MDI on retention, expressed in percentage point changes in retention, ranges from 4.9 to 6.6 percentage points, depending on the prevailing retention rate in the control condition (Table IV.3). Because outcome changes can be more reliably detected when starting from a more homogeneous outcome distribution, MDIs are lower to the extent that retention outcomes are more homogenous in the control group—that is, to the extent that a larger proportion of control group teachers are retained. Based on national statistics, we expect the control group retention rate to range from 80 to 90 percent (Ingersoll 2003), which determines the range of MDIs shown in Table IV.3. Within the full sample of teachers, those from tested grades and subjects are expected to outnumber those from nontested grades and subjects by more than two to one, and thus MDIs for the former subgroup are considerably smaller than those for the latter.

The MDIs on teacher retention are large relative to the fraction of control group teachers who leave; indeed, for a control group nonretention rate of 20 percent (represented by the last column of Table IV.3), the MDI of 6.6 percentage points amounts to a one-third reduction in the nonretention rate. However, the MDIs are well below the impact magnitudes that would make a policy-relevant contribution to raising student test scores. As discussed in section D, impacts on retention contribute to increased test scores only to the extent that leavers are more effective than entrants. Even if a 6.6 percentage point retention effect were scaled (multiplied) by a large leaver-entrant effectiveness gap of 0.3 SD of test scores—nearly twice the within-school standard deviation of

[14] To fully mimic the scenario in which student-level pretests are excluded, we also assume a (lower) student-level R-square of 0.2, but the student-level R-square has little influence on the MDI.

Table IV.2. Minimum Detectable Impacts on Student Test Scores Under Alternative Scenarios

Scenario	R-Square (School-Level / Student-Level)	ICC	Proportion of Students with Missing End-of-Year Scores	Proportion of Schools with Data Missing for All Students	Minimum Detectable Impact (in Standard Deviations of Student Test Scores)
Benchmark assumptions	0.65/0.4[a]	0.13	0.15	0	0.09
Low R-square	0.4/0.2	0.13	0.15	0	0.10
Very low R-square	0.2/0.2	0.13	0.15	0	0.12
High ICC	0.65/0.4[a]	0.16	0.15	0	0.10
Very high ICC	0.65/0.4[a]	0.20	0.15	0	0.11
High proportion of students with missing end-of-year test scores	0.65/0.4[a]	0.13	0.25	0	0.09
Very high proportion of students with missing end-of-year test scores	0.65/0.4[a]	0.13	0.35	0	0.09
High proportion of schools in which data are entirely missing	0.65/0.4[a]	0.13	0.15	0.2	0.10
Very high proportion of schools in which data are entirely missing	0.65/0.4[a]	0.13	0.15	0.4	0.11

Note: Assumptions not shown in this table are identical to those in Table IV.1.

[a]The R-square values shown pertain to middle schools. See the notes to Table IV.1 for benchmark R-square values pertaining to other school types.

ICC = intraclass correlation

Table IV.3. Minimum Detectable Impacts on Teacher Retention

Type of Sample	Number of Schools	Number of Teachers	Minimum Detectable Impact (in Percentage Points) if Control Group Retention Rate is: 90 Percent	80 Percent
Full sample	250	1,900	4.9	6.6
Teachers in tested grade-subject combinations	250	1,300	5.7	7.6
Teachers in nontested grade-subject combinations	250	600	7.9	10.6

Note: The calculations are based on the following assumptions: 80 percent power and a 5 percent significance level for a two-tailed test; 60 percent of schools in the sample will be elementary schools, 20 percent will be middle schools, and 20 percent will be K-8 or K-12 schools; the teacher sample will consist of 6 sampled teachers per elementary school, 8 per middle school, and 12 per K-8 or K-12 school; nonresponse rate in the teacher survey will be 15 percent; 6 percent of the total variance of retention outcomes will be between schools; covariates will explain none of the variance in retention outcomes.

teacher effectiveness—the resulting rise in student test scores would be only 0.02 SD. Thus, our study can detect the full range of retention effects that are likely to be relevant as mechanisms for influencing student achievement.

Other intermediate impacts for which we aim to conduct reasonably precise estimates are the impacts on the stayer-leaver and entrant-stayer effectiveness gaps. Although these impacts are estimated from variants of equation (1), the MDIs on these effectiveness gaps are larger than the MDIs on student test scores. Impacts on effectiveness gaps involve a double comparison—assessing the treatment-control difference in the test score gap between students of two specified groups of teachers—whereas impacts on test scores involve a single comparison of average outcomes between the treatment and control groups; more comparisons are accompanied with greater imprecision. On each type of effectiveness gap, we will be able to detect an (unscaled) impact of 0.10 SD based on administrative data—in which student test score records are linked with administrative records on teachers—or 0.12 SD based on teacher survey data linked with test scores in one elementary grade and one middle grade.[15] Upon being scaled by a 20 percent nonretention rate, these MDIs would account for a rise of 0.02 SD of test scores—again, well below the range of policy-relevant changes in student achievement.

F. Correlational Analyses

In light of the variation in TIF programs proposed by the grantees, we will examine whether districts with particularly strong or weak impacts share any distinctive approaches to performance pay. Because this study does not randomly assign schools (or grantees) to different program features, the study design cannot yield causal estimates of the effects of particular program features. In other words, outcome differences between districts could be due to the differences in program characteristics, but they could also be due to any number of other factors. We will report, however, on associations between program features and impacts. Although these associations cannot reveal causal relationships, they can suggest hypotheses regarding the elements of performance pay approaches that might contribute to impacts on key outcomes.

Meaningful variation in impacts across districts is necessary to detect any association between impacts and program features. After estimating the district-specific impacts on the basis of equation (1), we will conduct an F-test of the null hypothesis that all district-specific impacts are equal. If the variation in impacts is statistically significant, then we will explore the potential correlates of this variation.

Our approach to describing the association between program features and impacts is to estimate a district-level regression in which we model the estimated impact in district k, β_k, as a linear function of a specified performance pay feature, W_k:

[15] This calculation incorporates all of the assumptions from Table IV.1, except that if the teacher data to which student test scores are linked comes from the teacher survey, then the estimation sample includes only 80 students per elementary school, 260 students per middle school, and 50 students per K-8 or K-12 school. Additional assumptions include: (1) 80 percent of students in the sample are taught by stayers and 20 percent are taught by end-of-1 leavers (or, equivalently, start-of-2 entrants); and (2) the correlation between the school-level average effectiveness of stayers and leavers (or, equivalently, of stayers and entrants) is 0.5.

(4) $\hat{\beta}_k = \pi W_k + \nu_k$,

where ν_k is an error term that includes the error in estimating the district-specific impact. Because impacts might be more precisely estimated in some districts than in others, we will weight districts by the precision of the estimated impacts when estimating equation (4) to account for this source of heteroskedasticity in the error term.

Although the grantees are still in the process of finalizing the design of their TIF programs, we anticipate that various performance pay features might be quantifiable into uniformly defined measures whose association with impacts can be examined. These program dimensions include (1) the weight on individual versus group performance in awarding performance-based compensation; (2) the average and maximum compensation amounts that educators can receive from the performance pay component of the TIF program; and (3) the degree to which these payouts vary across educators. For each considered feature, we will estimate equation (4) with the specified program feature as the only covariate, given the limited number of districts in the sample.

G. Nonexperimental Comparisons of TIF to Non- TIF Schools

The FY 2010 TIF program requires grantees to implement a bundle of pay reforms for educators, including performance-based pay and other incentives and supports. The experimental impact evaluation described above will focus specifically on the performance-based pay, but policymakers and program planners are also eager to learn about the other incentives and supports. Here we consider a nonexperimental analysis that can shed light on this full package of supports.

1. Descriptive Analyses

We will calculate summary statistics of student achievement in TIF and non-TIF schools both before and during the intervention period. Because the outcomes of non-TIF schools are intended as points of reference for the outcomes of TIF schools, the two sets of schools should use comparable outcome measures—that is, administer the same student assessments. The set of non-TIF schools in this analysis will therefore consist of all non-TIF schools in the same states and grades as TIF schools.

2. Comparisons with "Similar" Non-TIF Schools

We will compare the outcomes of TIF schools with those of carefully selected groups of non-TIF schools. The non-TIF schools are intended to provide an informative benchmark against which to compare outcomes of TIF schools, but we do not intend to interpret the differences between TIF and non-TIF schools as being attributable solely to the TIF program. Unlike with the experimental analysis, several other factors, both observable and unobservable, may be confounded with TIF status.

We will select non-TIF comparison groups whose pre-intervention outcome trajectories and characteristics are as similar as possible to those of the TIF sample. This can be done using school-level data on test scores and other characteristics over time. In particular, we will attempt to create non-TIF comparison groups from schools within the same states as TIF schools because they need to share the same assessments and state policy environments. Furthermore, if a TIF district includes

a sufficient number of similar non-TIF schools, we will construct the comparison group from schools within the same district to account for similar district environments.

3. Sample and Data Requirements

We expect to focus on TIF schools that are already participating in the national evaluation. This approach will ensure that we already have access to data and understanding of the local context. Within the set of TIF districts, a number of criteria determine whether a given district is suitable for the nonexperimental analysis. First, the TIF district must not have implemented another major reform near the time that it first implements TIF-funded reforms; otherwise, outcome trends for this district may not accurately indicate the likely trends for districts that implement only a TIF-funded compensation reform package. Second, in order for us to identify a useful comparison sample for a TIF district, the district must be located in a state with available test score data and an ample pool of non-TIF districts that have not implemented similar compensation reforms. For the same reason, the TIF district cannot be unique within the state in terms of characteristics or pre-intervention outcome trajectories (or if it is, then schools in that district must be comparable to schools in other parts of the state).

To carry out this approach in a given state, we will seek annual school-level data on test scores and demographic characteristics for all schools in the state. The data must cover several years before the 2010–2011 school year so that we can select comparison samples with similar pre-intervention outcome trajectories as TIF samples. If we create comparison samples through school-level matching, the data must be at the school level.

An important challenge in this analysis is to describe the policy environment of the comparison samples and to select, to the extent possible, comparison samples with no major programs resembling TIF reforms. Our proposed approach is to collect public information on districts that participate in large, well-known programs involving compensation reforms, including Race to the Top, Gates Intensive Partnerships, and major state initiatives. We can then evaluate on a case-by-case basis whether to exclude these districts from our comparison samples.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics*, vol. 25, no. 1, 2007, pp. 95–135.

Angrist, Joshua, Guido Imbens, and Donald Rubin. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* , vol. 91, 1996, pp. 444-472.

Bayonas, Holli. "Guilford County Schools Mission Possible Program: Year 3 (2008–09) External Evaluation Report." Greensboro, NC: The SERVE Center, University of North Carolina at Greensboro, 2010.

Bloom, Howard, Carolyn Hill, Alison Black, and Mark Lipsey. "Performance Trajectories and Performance Gaps as Achievement Effect Size Benchmarks for Educational Interventions." MDRC Working Papers on Research Methodology. New York, NY: MDRC, October 2008.

Bloom, Howard, Lashawn Richburg-Hayes, and Alison Black. "Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions." *Educational Evaluation and Policy Analysis*, vol. 29, no. 1, 2007, pp. 30–59.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah Rockoff, and James Wyckoff. "The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-Poverty Schools." *Journal of Policy Analysis and Management*, vol. 27, no. 4, 2008, pp. 793–818.

Deke, John, Lisa Dragoset, and Ravaris Moore. "Precision Gains from Publically Available School Proficiency Measures Compared to Study-Collected Test Scores in Education Cluster-Randomized Trials." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, October 2010.

Fryer, R. (2011). *Teacher incentives and student achievement: Evidence from New York City public schools.* Cambridge, MA: National Bureau of Economic Research.

Glazerman, Steven. "Teacher Compensation Reform: Promising Strategies and Feasible Methods to Rigorously Study Them." Washington, DC: Mathematica Policy Research, January 2004.

Glazerman, Steven, Allison McKie, and Nancy Carey. "An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year One Impact Report." Washington, DC: Mathematica Policy Research, April 2009.

Glazerman, Steven, and Allison Seifullah. "An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year Two Impact Report." Washington, DC: Mathematica Policy Research, May 2010.

Glazerman, Steven, Eric Isenberg, Sarah Dolfin, Martha Bleeker, Amy Johnson, Mary Grider, and Matthew Jacobus. "Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, June 2010.

Glazerman, Steven, and Jeffrey Max. "Mapping the Prevalence of High-Performing Teachers." Washington, DC: Mathematica Policy Research, 2011.

Goodman, Sarena, and Lesley Turner. "Does Whole School Performance Pay Improve Student Learning? Evidence from the New York City Schools." *Education Next*, vol. 11, no. 2, Spring 2011.

Gordon, Robert, Thomas Kane, and Douglas Staiger. "Identifying Effective Teachers Using Performance on the Job." Hamilton Project Discussion Paper 2006-01. Washington, DC: The Brookings Institution, 2006.

Hanushek, Eric. 2010. "The Difference is Teacher Quality." In *Waiting for "Superman": How We Can Save America's Failing Public Schools*, edited by Karl Weber. New York: Public Affairs, 2010, pp. 81-100.

Hanushek, Eric, John Kain, Daniel O'Brien, and Steven Rivkin. "The Market for Teacher Quality." NBER Working Paper 11154. Cambridge, MA: National Bureau of Economic Research, February 2005.

Hanushek, Eric, and Steven Rivkin. "Generalizations About Using Value-Added Measures of Teacher Quality." *American Economic Review (AEA Papers and Proceedings)*, vol. 100. no. 2, 2010, pp. 267–271.

Hanushek, Eric, and Steven Rivkin. "Teacher Quality." In *Handbook of the Economics of Education,* edited by Eric Hanushek and Finis Welch. Amsterdam: North-Holland, 2006.

Harris, Douglas, and Tim Sass. "Teacher Training, Teacher Quality, and Student Achievement." Unpublished. Tallahassee: Florida State University, 2008.

Hatry, Harry P., John M. Greiner, and Brenda G. Ashford. *Issues and Case Studies in Teacher Incentive Plans*. Second edition. Washington, DC: The Urban Institute Press, 1994.

Iatarola, Patrice, and Leanna Stiefel. "Intradistrict Equity of Public Education Resources and Performance." *Economics of Education Review*, vol. 22, no. 1, 2003, pp. 69–78.

Ingersoll, Richard. "Is There Really a Teacher Shortage? A Research Report." Seattle, WA: Center for the Study of Teaching and Policy, University of Washington, 2003.

Jacob, Brian A. "The Effectiveness of Staffing Urban Schools with Effective Teachers." *The Future of Children*, vol. 17, no. 1, 2007, pp. 129–153.

Jacob, Brian A., and Steven Levitt. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics,* vol. 118, no. 3, 2003, pp. 843-877.

Jacob, Brian A. "Accountability, Incentives, and Behavior." *Journal of Public Economics*. vol. 89, no. 5-6, 2005, pp. 761-795.

Jordan, Heather, Robert Mendro, and Dash Weerasinghe. "Teacher Effects on Longitudinal Student Achievement." Paper presented at the CREATE Annual Meeting, Indianapolis, IN, July 1997.

Kane, Thomas, Jonah Rockoff, and Douglas Staiger. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review*, vol. 27, no. 6, 2008, pp. 615–631.

Kirby, Sheila, Scott Naftel, and Mark Berends. "Staffing At-Risk School Districts in Texas: Problems and Prospects." Santa Monica, CA: Rand Education, 1999.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., Epstein, S., Koppich, J., Peng, A. (2011). *A big apple for educators: New York City's experiment with schoolwide performance bonuses* (Final Evaluation Report). Santa Monica, CA: RAND Corporation.

Monk, David H. "Recruiting and Retaining High-Quality Teachers in Rural Areas." *The Future of Children*, vol. 17, no. 1, 2007, pp. 155–174.

Murnane, Richard, and David K. Cohen. "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive." *Harvard Educational Review*, vol. 56, no. 1, February 1986, pp. 1–17.

Podgursky, Michael. "The Single Salary Schedule for Teachers in K-12 Public Schools." Discussion paper prepared for the Center for Reform of School Systems. Columbia, MO: University of Missouri-Columbia, August 2002.

Podgursky, Michael, and Matthew Springer. "Teacher Performance Pay: A Review." *Journal of Policy Analysis and Management*, vol. 26, no. 4, 2007, pp. 909–949.

Rivkin, Steven, Eric Hanushek, and John Kain. "Teachers, Schools, and Academic Achievement." *Econometrica*, vol. 73, no. 2, 2005, pp. 417–458.

Rockoff, Jonah. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review (AEA Papers and Proceedings)*, vol. 94, no. 2, 2004, pp. 247–252.

Rubin, Donald B. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, vol. 66, 1974, pp. 688-701.

Sanders, William, and June Rivers. "Cumulative and Residual Effects of Teachers on Future Academic Achievement." Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center, November 1996.

Schacter, John, Tamara Schiff, Yeow Meng Thum, Cheryl Fagnano, Micheline Bendotti, Lew Solmon, Kimberly Firetag, and Lowell Milken. "The Impact of the Teacher Advancement Program on Student Achievement, Teacher Attitudes, and Job Satisfaction." Santa Monica, CA: Milken Family Foundation, 2002.

Schacter, John, Yeow Meng Thum, Daren Reifsneider, and Tamara Schiff. "The Teacher Advancement Program Report Two: Year Three Results from Arizona and Year One Results from South Carolina TAP Schools." Santa Monica, CA: Milken Family Foundation, 2004.

Silman, Timothy, and Steven Glazerman. "Teacher Bonuses for Extra Work: A Profile of Missouri's Career Ladder Program." Washington, DC: Mathematica Policy Research, May 2009.

Solmon, Lewis, J. Todd White, Donna Cohen, and Deborah Woo. "The Effectiveness of the Teacher Advancement Program." Santa Monica, CA: National Institute for Excellence in Teaching, 2007.

Springer, Matthew, Dale Ballou, and Art Peng. "Impact of the Teacher Advancement Program on Student Test Score Gains: Findings from an Independent Appraisal." Working Paper 2008-19. Nashville, TN: National Center on Performance Incentives, Vanderbilt University, 2008.

Springer, Matthew, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel McCaffrey, Matthew Pepper, and Brian Stecher. "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching." Nashville, TN: National Center on Performance Incentives, Vanderbilt University, 2010.

Springer, Matthew, Jessica Lewis, Michael Podgursky, Mark Ehlert, Lori Taylor, Omar Lopez, and Art (Xiao) Peng. "Governor's Educator Excellence Grant (GEEG) Program: Year Three Evaluation Report." Nashville, TN: National Center on Performance Incentives, 2009a.

Springer, Matthew, Jessica Lewis, Michael Podgursky, Mark Ehlert, Timothy Grownberg, Laura Hamilton, Dennis Jansen, Brian Stecher, Lori Taylor, Omar Lopez, and Art (Xiao) Peng. "Texas Educator Excellence Grant (TEEG) Program: Year Three Evaluation Report. Nashville, TN: National Center on Performance Incentives, 2009b.

Springer, Matthew, and Marcus Winters. "The NYC Teacher Pay-for-Performance Program: Early Evidence from a Randomized Trial." Civic Report No. 56. New York, NY: Manhattan Institute for Policy Research, 2009.

Tennessee Department of Education. "Tennessee's Most Effective Teachers: Are They Assigned to the Schools That Need Them Most?" Nashville, TN: Tennessee Department of Education, 2007.

MATHEMATICA
Policy Research

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC